

# A model of infant speech perception and learning

Philip Zurbuchen

September 16, 2016

# Acknowledgements

I'd like to thank Jochen Triesch for his supervision over the last year, during which he introduced me to the field of computational neuroscience. I thoroughly enjoyed his lectures (for instance on Reinforcement Learning) and his insights and advice whenever we discussed my work.

Many thanks to Max Murakami. Having (in his master thesis) developed the original speech acquisition model, Max did a great job of patiently lecturing me on the ins and outs of *Listen and Babble* and its context: speech acquisition in the human brain. I learned a lot this past year in many areas; much of that is attributed to him.

I thank my mother, Sarah Zurbuchen, for turning Swiss-English into proper English over the years, but also for correcting remaining spelling errors in this thesis.

Special thanks to my dear wife, Tirza, who was willing to put up with my long programming evenings and politely listened to my musings about how 'this-and-that' might be connected to 'that-and-that' (before changing my mind again). She is a social worker and I'm glad she can cope with a case like me!

Lastly, as did greater men like Newton, Maxwell and Faraday, I dedicate my efforts to the Creator of speech, who himself spoke through Christ Jesus.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Infants learn speech . . . . .	1
1.1.1	The challenge of speech perception . . . . .	1
1.1.2	The challenge of speech imitation . . . . .	5
1.1.3	Infant-directed speech . . . . .	8
1.2	Listen and Babble – A model of vowel acquisition through imitation	12
1.2.1	Basic concepts . . . . .	12
1.2.2	The model . . . . .	13
1.2.3	Some results . . . . .	15
1.2.4	Open questions . . . . .	16
<b>2</b>	<b>Methods</b>	<b>18</b>
2.1	VocalTractLab . . . . .	18
2.2	Auditory System . . . . .	21
2.3	RL algorithm . . . . .	24
<b>3</b>	<b>Results</b>	<b>25</b>
3.1	Ambient speech . . . . .	25
3.1.1	Age-specific pitch . . . . .	25
3.1.2	Speaker series . . . . .	25
3.1.3	Vowel shape settings . . . . .	25
3.1.4	Vowel Samples . . . . .	29
3.2	Training reservoirs . . . . .	32
3.2.1	Partial training . . . . .	32
3.2.2	Comparing reservoir training paradigms . . . . .	33
3.3	Imitation . . . . .	36
3.4	Motherese . . . . .	38
<b>4</b>	<b>Discussion</b>	<b>41</b>
<b>5</b>	<b>Listen and Babble with Caregiver Imitation</b>	<b>45</b>
<b>6</b>	<b>For developers</b>	<b>49</b>
6.1	Getting started . . . . .	49
6.2	Project structure . . . . .	49

# 1 Introduction

## 1.1 Infants learn speech

You learned to speak as an infant. We all did – we learned speech without knowledge or concept of the very language in which we now so naturally articulate our thoughts, feelings and questions. This in itself is one of the many amazing facts about this universe.

In my thesis, I introduce the reader to the learning process involved in acquiring speech. We'll focus on specific challenges involved in learning to speak a language from scratch. And while we're at it, I'll point out a few important aspects of speech itself. How is speech produced in a human being, and what is man's instrument when he says "I love a good conversation!" ?

Then in section 1.2, we'll review previous work done by M. Murakami in 2014: the original *Listen and Babble* model. My work is complementary to *Listen and Babble* in that I seek to discuss and extend the model. Restating all the details of Murakami's work would go too far – I refer to his thesis and publication for more detail [1, 2]. I shall cover some important points already mentioned in Murakami's work, specifically concerning the shortcomings of the model and ideas on improving parts of the model. My research addresses some of these shortcomings and seeks to improve the conceptual framework of *Listen and Babble* by proposing new ways of thinking about such issues.

But first, we must start at the beginning and ask the question: when do infants start learning to speak a language?

### 1.1.1 The challenge of speech perception

We started to learn speech very early, even before our first utterances. Research shows that infants start mapping critical aspects of ambient speech in the first year of life [4]. Ambient language sounds are listened to and analysed before ever understanding a word [5, 6].

Infants are confronted with lots of challenges when trying to hear and recognise speech sounds. Of course the infant must learn to perceive speech well in order to start loading syllables and words with meaning, or even imitating the caregiver's speech.

**Recognising speech sounds** We are quick to think that infants hear well defined and clear language signals. We think they "hear what we say" and merely have to learn the meaning of the words – perhaps a bit like a code breaker, trying to make sense of a jumble of (seemingly randomly arranged) typed out letters. But actually the task of hearing language is much harder. Infants have to learn speech *perception* first.

This becomes more clear when we are presented with examples from adult foreign-language perception. Native German speakers have problems hearing the difference between English speech sounds [ð]<sup>1</sup> (as in **this**) and [θ] (**thing**).

---

<sup>1</sup>Phones/Phonemes given in IPA (International Phonetic Alphabet) notation. See appendix for an overview of all IPA symbols.



While /ð/ is voiced, /θ/ is voiceless. Nonetheless, these both sound like the same speech sound for many non-English speakers<sup>2</sup>.

People who grow up with different ambient language perceive speech differently. An infant must learn perception of his ambient language as a first step. It is easy to see that the example of the code breaker, who reads well defined, typed out alphabetical symbols, does not go far enough. Instead, the infant code breaker is confronted with a continuous signal, without any gaps between letters (or words!).

When listening to our native language, we think we hear separate words when spoken. But this is not the case. I recorded myself saying “I like apple pie” and plotted waveform and spectrogram (see Box 1). Confronted with the physical signal, finding the end of one word and the beginning of the next is not trivial. Also, notice that we all would hear two p-sounds in the sentence “I like **apple pie**”. But finding the p-sounds in the spectrogram is not that easy. And when we’ve found them, we realise that they aren’t the same at all, due to their *phonetic context* (i.e. which sounds come before or after).

**Phones and phonemes** This calls for a way of describing language sounds on an abstract level. We all hear language-specifically, which means we group *phones* (speech sounds) into abstract speech units called *phonemes*. A phoneme is “the smallest unit of speech that can be used to make one word different from another word” [9]. Different phones are perceived to be realisations of one phoneme if they fulfill the same function in a language.

In order to understand the difference between the concept of phones and phonemes, consider this illustration; All devoted mountaineers in my home-country (Switzerland) have come across their fair share of Edelweiß<sup>3</sup> sightings. Notice that whenever one is sighted (lazier readers will simply google “Edelweiß”), we perceive a *realisation* of Edelweiß. The word “Edelweiß” is our class into which we group sightings of flowers with specific features (white petals, yellow inflorescences in the middle). This perceptual class is abstract and stands for a large variety of realisations in nature (some small, some large, some more fluffy, some smoother). Similarly, phonemes are abstract units into which we group actual speech sounds (phones).

Boundaries between phonemes are language specific. Take, for example, the German **ich**- and **ach**-sound ([ç] and [x] in the phonetic alphabet). These are two physically distinguishable phones (they are also produced in different ways), but they are generally perceived as the same language building block (phoneme) of the German language. Or take the vowels [e] and [i], which are clearly two

---

<sup>2</sup>In fact, even English-speaking children have difficulties in distinguishing between /ð/ and /θ/ (which are among the last phonemes learned in the English language). These are often confused by young children – frequently even until they start elementary school.

“Prior to this age [5 years], many children substitute the sounds [f] and [v] respectively. For small children, *fought* and *thought* are therefore homophones [phones which are perceived as the same phoneme]. As British and American children begin school at age four and five respectively, this means that many are learning to read and write before they have sorted out these sounds, and the infantile pronunciation is frequently reflected in their spelling errors: *ve fing for the thing*.” [10]

<sup>3</sup>Alpine Edelweiß (*Leontopodium nivale*) is a white flower, often found in alpine regions. They grow in rock cracks and meadows at high altitudes.

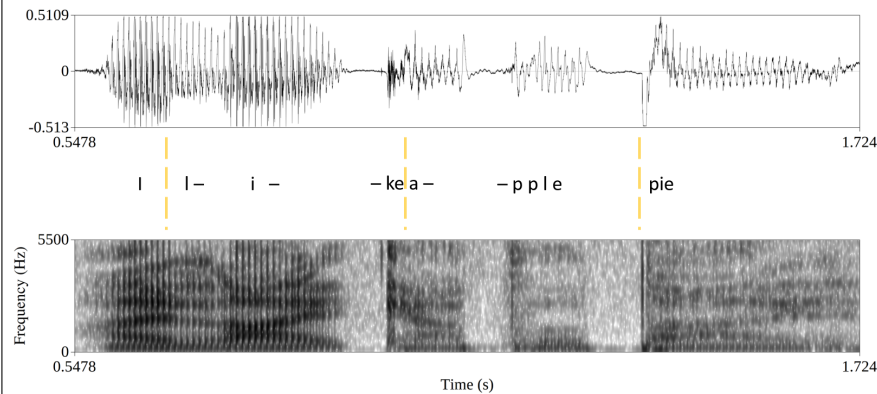
## Waveform and Spectrogram

## Box 1

We can analyse sound using either its waveform or spectrogram. A waveform is simply the pressure wave of the speech sound (which arrived at the microphone/human ear) plotted over time. The waveform can be broken down into all the single frequencies that contribute to the waveform. The power in each of these frequencies (*spectral density*) can be plotted over time, too. A plot of the spectral density over time is called a spectrogram.

We do well to look at spectrograms when thinking about speech input. Sound, after arriving in form of mechanical waves in the inner ear, is transformed in the cochlea into neural spiking. Each neuron location in the cochlea responds to a different frequency. This tonotopic organisation (spatial layout of frequencies) is repeated in other areas of the brain (inferior colliculus, primary auditory cortex). The brain effectively works with something like a spectrogram of the sound input. Spectral density over time is therefore a suitable representation of speech signals.

Frequencies at spectral power maxima (darkest colouring in the spectrogram below) are called formant frequencies ( $F_1, F_2, F_3$ ). The maximum which is lowest in frequency is called the pitch, or  $F_0$ . We can think of the pitch as a measure of “how high a sound is”. If I sing the musical note  $C_1$  for example, I’ll have an  $F_0$  of 32.70 Hz.



(a) Waveform (above) and spectrogram (below) of the sentence “I like apple pie”. In the spectrogram, darker colour denotes higher spectral density. Word borders are marked by vertical dashed lines. Extracted and drawn using Praat [8].

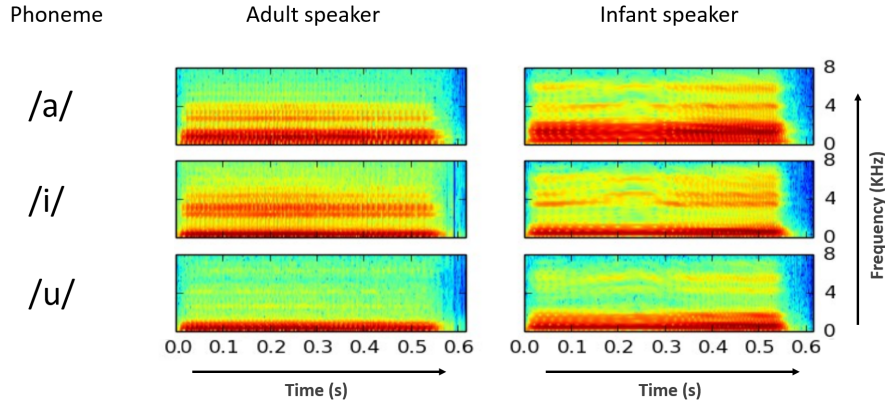


Figure 1: Acoustic spectrograms of 3 different vowels ( $/a/$ ,  $/i/$ ,  $/u/$  from top to bottom) from 2 different speakers (left: adult, right: infant). Colour denotes spectral power (red: high, blue: low).

different phones (and – in germanic languages – two different phonemes). Yet many native Chinese speakers have difficulties hearing the difference between the two.

In any specific language, phones within one phoneme are meant to be perceived as representing the same language building block. Infants must learn the building blocks of their ambient language before they can meaningfully put blocks (phonemes) together to hear words. Thinking back to the code breaker in the last paragraph, one might say: the codebreaker must first learn to decipher the bad handwriting, and only then can he start making sense of the (now clearly perceived but still encoded) symbols of the alphabet.

**Generalising across speakers** The variety among phones belonging to one phoneme is especially striking when we consider speech from different speakers. In Figure 1 we see acoustic spectrograms of three German vowels  $/a/$ ,  $/u/$  and  $/i/$ . Note that a spectrogram reflects what we actually physically receive in terms of sensory information when we hear a sound. While spectrograms from different vowels do differ (moving up and down between spectrograms in Figure 1), so do spectrograms of the same vowel spoken by different speakers (moving left and right)! In this case, an adult speaker is compared to an infant speaker when pronouncing the German vowels  $/a/$ ,  $/i/$  and  $/u/$ <sup>4</sup>.

This is called the problem of *speaker normalisation* [12] (also sometimes coined *speaker generalisation*). Infants learn to identify phones that are physically different, but belong to the same phoneme. “If my 4-year-old sister says X, it’s the same thing as when my Grandpa says Y.” – X and Y sounding very different, but, in that specific language, perceived as one and the same phoneme.

**Further difficulties** Even when uttered by the same speaker, phonemes will have a large variety of physical realisations. For instance, phones of one phoneme

<sup>4</sup>Vowel sounds were produced using the adult and the infant speaker in the speech synthesizer *Vocaltractlab* [11].

will vary a lot depending on the *phonetic context*, i.e. where a phone is placed in a word [13]. Also, slow speech results in different acoustic properties than fast speech [14]. Physical features of phones are often highly overlapping as we will see later on in section 3.1.

For all vocal gestures in Figure 1 there was not much pitch alteration (only approx. half a semitone). Also, the vowels were pronounced alone, without other phonemes before or after. The reader can well imagine that adding language melody (varying pitch) and phonetic context to these vowels would make them even less recognisable.

State of the art computer speech recognition software cannot recognise phones as belonging to the same class when the speaker, or the rate of speech or the phonetic context changes [15]. And yet, infants are shown to perceive phonetic classes over all these variabilities [13, 14, 16].

**Encoding phonetic features** So how does the human brain manage to generalise so well? Recent work has been done measuring direct cortical activity from patients undergoing clinical evaluation for epilepsy surgery. Electrode arrays were implanted on the superior temporal gyrus (STG), which is attributed to language perception. The recordings showed phonetic feature encoding, – Neuron clusters in STG responded to spectrotemporal acoustic cues of the speech sounds [17]. One of these spectrotemporal acoustic cues is frequently used throughout this thesis: vowel formants in formant space (see Box 2).

However, such findings underline the complexity of the task of language perception and the fact the infants (who live up to the challenge) seem to be natural language-learners in a way that is not yet understood.

We conclude that, in order to learn speech, an infant must first learn to cluster phones into groups of phonemes, which in turn fulfill a specific function in that language. A newly-born is thought to listen to ambient speech and train its own auditory system to do this “categorisation” correctly. Only then, eventually, can a sequence of speech sounds be loaded with actual meaning. And, moving on, knowing how phones *should* sound is crucial for starting to imitate them.

### 1.1.2 The challenge of speech imitation

We have seen some challenges that infants face when learning to perceive language throughout the previous section. In general, getting your perception right is crucial for imitating correctly. This is the underlying assumption I make throughout my thesis. We will, however, discuss this point again in Section 5.

The afore mentioned difficulties involved in the perceptual aspect come along with the staggering challenge of actually imitating speech sound. We understand how difficult playing the violin is, not by hearing a violin solo, but by investigating sound production on a violin itself. In this section I’d like to introduce the ‘instrument’ by which we produce speech and then go on to a phase in child development called ‘babbling’, in which infants are thought to learn to control this speech instrument – their own vocal tract.

**The Vocal Tract** The vocal tract is comprised of all parts actively involved in speech production. Figure 2 illustrates the complexity of the human vocal tract.

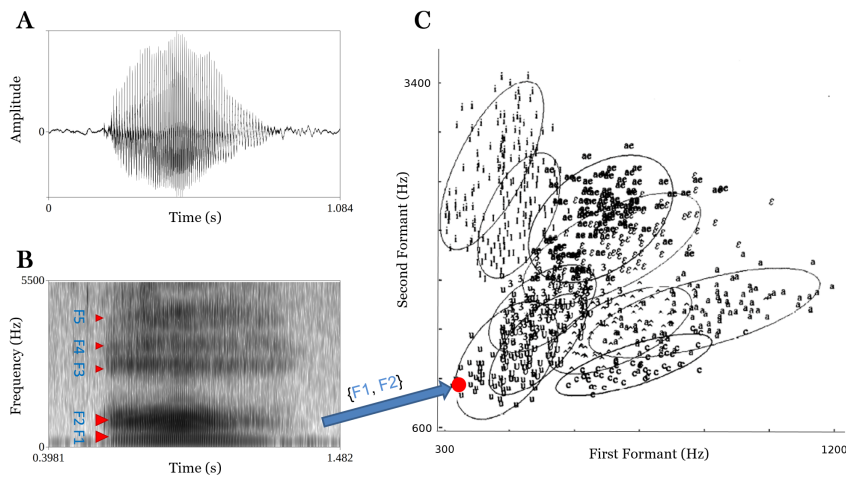
## Formant space

## Box 2

### *and phonetic features of vowels*

A concept frequently used throughout this thesis is formant space. Below, the German vowel /a/ is shown as a waveform (**A**) and its corresponding spectrogram (**B**). Prominent (and stable) maxima of the spectrum of a sound are called formants. Vowels are comparable to musical chords in that each vowel has a set of formants which together make up its characteristic sound.

In the spectrogram of the German vowel /a/, the formant frequencies are shown (without showing the pitch, which is hard to distinguish in this plot). Taking the first two formants ( $F_1$  and  $F_2$ ) we can diametrically plot them in a space called formant space. This specific realisation of the German vowel /a/ (author's pronunciation) is drawn in formant space (**C**, red arrow) along with formants from a study of American vowels [25]. (Notice, the American 'short u' vowel is similar to the German vowel /a/.)



A and B: Time waveform and spectrogram of the German vowel /a/ (produced by the author). C: American vowels from many different speakers plotted in formant space. (Source: [25])

We see that, although formant space is a practical way of showing (somewhat-) distinguishing vowel properties, vowels are all but clearly defined. Groups of different vowel formants strongly overlap. Keeping vowels apart by looking at their physical features (e.g. formants) is not simple.

Additionally, measuring formants is not at all trivial [27]. In this thesis all formants were evaluated using the Burg method [26].

Whenever one formant value is shown for a certain phone from a certain speaker, the medium was taken over all formant values which were extracted from the vowel over time.

In order to produce speech, approx. 100 muscles are simultaneously involved [28]. Using this highly complex subsystem to articulate a phrase would correspond to a 100-dimensional trajectory, where a certain motor configuration would have to develop with precise timing (thus a trajectory). The figure shows various layers of muscles:

- thorax and back muscles, which are important for lung contraction and airflow production,
- throat and neck muscles. These include the muscles around and between the cartilage that makes up the larynx. In the larynx ('voicebox') the vocal folds are stretched/relaxed, and the glottis (the opening between the vocal folds) is made wider/narrower. When air passes the (nearly closed-) vocal folds they begin to vibrate, thus creating a tone which is then 'shaped' in the rest of the vocal tract,
- craniofacial muscles, together with the tongue, are mainly responsible for articulation of speech sounds.

In principle, every one of these muscles has to be controlled by the human brain. However, we can simplify the problem by neglecting most of these muscles and focusing on only a few. In phonetics, it is common to look at a handful of *articulators*, which include the lips, teeth, alveolar ridge, hard- and soft palate, pharynx and larynx (see Figure 8, left). The tongue itself is often labeled according to its segments: The tip, blade, front, back or root of the tongue.

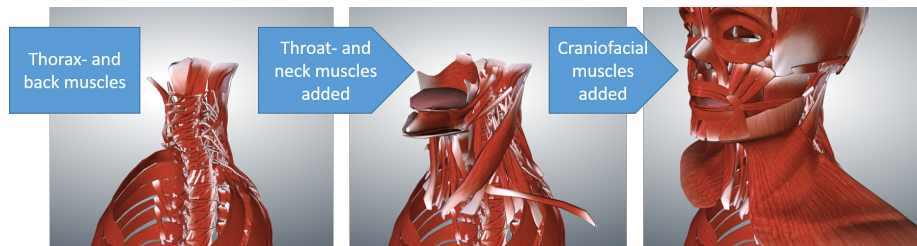


Figure 2: The muscles of the human vocal tract, shown in different layers. Image created using software freely available [22].

**Infant babbling** Having given an impression of our 'voice organ', the instrument that has to be explored and learned by the infant learner, we now move on to consider *how* the infant learns to articulate speech with his own vocal tract.

Infants seem to explore their own capability to produce vowels (cooing, around 3 months) and strings of simple syllables (around 7 months) [29]. This process, which goes on until they are able to produce recognisable words (normally at around 12 months) is called *babbling*. Strings of syllables produced around 7 months often include small alterations over time (for example "baba", "bebe"). This specific phase is generally known as *canonical babbling*. Babbling seems to be present in all children acquiring language and has been studied in infants across the world. An intriguing aspect of babbling is that it does not need to be connected to the infant's emotional state. Infants seem to babble spontaneously and incessantly when they are both emotionally calm or upset/excited [30].

The universal nature of infant babbling might suggest that infants use exploration (of their own vocal tract's "states") and, reinforcing those lucky vocal gestures which actually sounded 'right', actually begin to imitate their caregivers. Research shows that, when babbling, infants do match patterns of rhythm of the language(s) to which they are exposed [31]. They use intonation patterns (as in Box 3) and timing of surrounding speech, mainly using the consonants and vowels that occur most frequently in the caregiver's language [31].

It is important to see that babbling is not the only stage important for language learning. In Figure 4 canonical babbling is represented as one stage in the context of a whole range of production- (and perceptual) stages. Some of these activities seem to happen simultaneously. Humans seem to be equipped with a diverse set of abilities which, combined and rightly timed, on the perceptual side as well as on the production side, contribute to the acquisition of language and speech.

### 1.1.3 Infant-directed speech

**Motherese** Another source of help for the young learner is deliberate infant-caregiver interaction. When we talk to infants and young children, we use a special 'mode' of speech called *motherese*. Caregivers in most countries use motherese when addressing children. The fact that motherese is a common phenomenon suggests that infants need infant-directed speech to learn language. When addressing infants, we tend to express ourselves with a lot of change in pitch (see Figure 5 a). Also, motherese is typically slower and better articulated. An important aspect in vowel pronunciation is the larger vowel space spanned by the formants when infant directed (see Figure 5 b). Studies among English, Russian and Swedish mothers show that the medium formants of their vowels are further apart (e.g. /u/ has slightly lower  $F_1$  and  $F_2$  values while /i/ has lower  $F_1$  but higher  $F_2$  values for motherese speech, compared to adult directed speech).

In Figure 5 b triangles are drawn by connecting positions of /u/, /i/ and /a/ in formant space. Those three vowels act as extreme vowel sounds. Other vowels are articulated somewhere inside the triangle spanned by /u/, /i/ and /a/. This is why phoneticists use the concept of vowel triangles. Larger triangles mean that all vowels (the corner points /u/, /i/ and /a/, but also those inside the triangle – /e/ and /o/ for example) will be further apart and more easily distinguishable.

**Caregivers imitate** Infant directed speech (i.e. motherese) is thought to aid the young learner in the perceptual task. But recent research suggests that parent-child interaction might also directly contribute to the learning process (not merely on the perceptual side)<sup>5</sup>. We naturally think of imitation being done by the infant (infants imitate their parents). But mothers frequently imitate their infants' speech – in fact more often than the other way round (e.g. [35], and other studies quoted in [52]). This imitation might form associations between the learner's speech gesture and the caregiver's imitatory response. The infant learner might, for example, execute a certain motor pattern which (by

<sup>5</sup>See "Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation", section 4 (Support for Mirrored Equivalence). Full citation: [52]

### Are young infants imitation-driven?

### Box 3

The subject of imitation is debated among child researchers. It is not yet clear whether actions of infants in their first months are truly motivated by the desire to imitate or not. Studies on infants younger than 3 weeks showed that these sometimes do match adult behaviour, but to some it seems unlikely that this matching is imitation-driven [18].

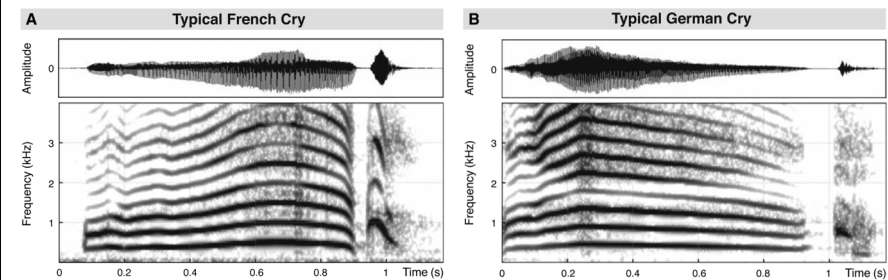
Studies on tongue protrusion (among other facial gestures) showed above-baseline activity, when the caregiver also stuck out his/her tongue (e.g. [19]). But is increased tongue protrusion the result of imitation, or simply of the infant's arousal when it sees something interesting (the caregiver sticking out his/her tongue)?

Recent meta-analysis shows poor statistical power for some studies that speak *against* facial imitation and draws attention to the fact that in some cases the experimental setup choked the motivation to imitate [20].

We know that infants' attempts at producing speech strongly resemble ambient speech. Figure 3 shows typical French vs typical German infant crying. The ambient French speech intonation patterns being predominantly upward, the infants (who hear ambient speech from the 6th month of pregnancy on) seem to have the same upward intonation pattern when crying (and vice versa with German infants) [21].

Throughout this thesis, I assume infants are driven by the desire to imitate, also when learning speech. In the discussion section, I will compare this underlying assumption and look at models of speech acquisitions that take different stances.

Figure 3: Time waveform and spectrograms of a typical French cry and a typical German cry. Source: [21]





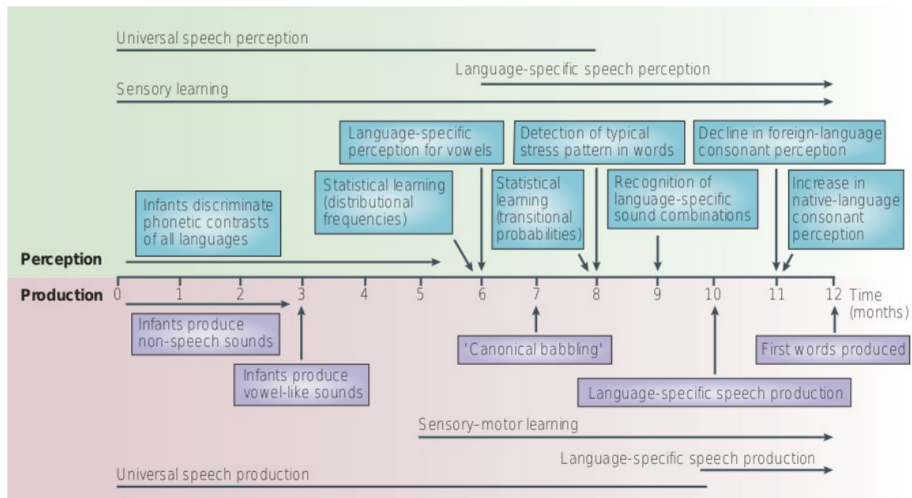


Figure 4: The development of speech perception and speech production in typically developing infants during their first year of life. Taken from [7].

chance) produces something similar to an actual speech sound. In response, the caregiver imitates and produces a correct pronunciation of that specific speech sound. In the process, the infant might form associations between the 'motor pattern' which was executed in order to perform the gesture and the auditory signal received from the caregiver (the correct pronunciation). This motor pattern would then be reinforced and used again (maybe with a slight variation / exploration).

For now let it be said that the role of the caregiver, specifically when using infant directed speech, seems crucial to language acquisition of infants. But *Listen and Babble* works without direct caregiver-infant interaction while the model learns to articulate. I argue that this could be the reason for certain problems *Listen and Babble* faces.

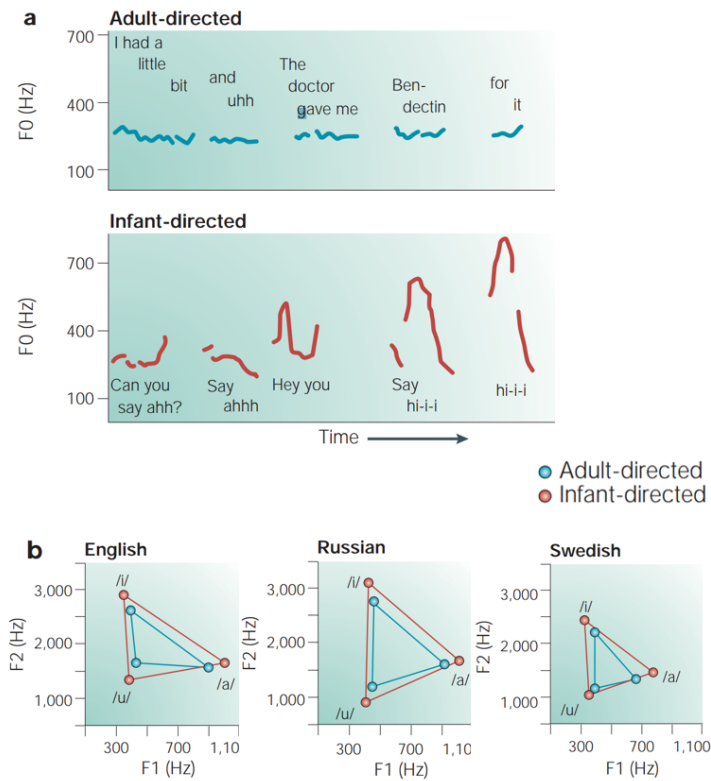


Figure 5: **a**: Characteristic pitch contours of adult-directed speech and motherese (*see caption*). Large pitch variation is typical for infant-directed speech. **b**: Vowel triangles for adult-directed and motherese speech in formant space from studies in three different languages. All three studies show increased vowel-space for motherese speech (Vowels are, formant wise, more distinguishable). Taken from [7].

## 1.2 Listen and Babble – A model of vowel acquisition through imitation

My research builds upon Murakami’s work, described in his thesis [2]. In short, Murakami successfully simulated vowel acquisition using machine learning. For a detailed description of his work I again refer to [2]. I shall describe the basics in this section, but also try to show some points where the model still needs improving.

### 1.2.1 Basic concepts

The model of vowel acquisition through imitation: *Listen and Babble* seeks to understand how speech is learned by infants. Our brains work with rewarding signals (e.g. dopamine<sup>6</sup>) and “use them for influencing brain activity that controls our actions, decisions and choices” [36]. Predictions are then made, based on rewards in the past, and prediction errors computed. These prediction errors then help to form new (and, over time, better) predictions for the future. These predictions are strongly linked to actions. In speech, for example, the prediction would be the actual sound produced by executing a combined position of the articulators. Like this, the infant would learn to predict the outcome (sound) of its own actions (vocal gesture) with steadily increasing precision. Muscle movements (or motor patterns in the cortex) are mapped to the infant’s own speech sounds: The young human learns to speak.

Before introducing the full model *Listen and Babble*, we must first cover some key concepts of the model.

**Machine learning** Learning new actions based on previous outcomes is central in a field called *machine learning*. Machine learning seeks to produce increasingly ‘correct’ actions from a machine (e.g. a programmable robot) without constantly giving explicit instructions to the machine. There are various ways of doing this, each suitable for a certain range of applications. For now, let us look at two areas of machine learning<sup>7</sup>:

---

<sup>6</sup>To be precise: Dopamine codes for the so-called ‘prediction error’, the discrepancy between prediction and actual reward.

<sup>7</sup>For a more detailed (and valuable!) description of machine learning, explained in context of the *Listen and Babble* model, see [2], pages 12-20.

### Supervised Learning

Let's assume we are given a set of data. Each set can either belong to a category or produce some state. The task is to programme a computer to learn to make accurate predictions of the label (category) or state if a new data set is presented to it. In order to make good predictions, the programme has to be trained using sets of data that already are labeled, or where the produced state is known. The learner is 'supervised' in that each training set is pre-labeled. Only after training, when unlabeled datasets are presented, does the trained programme make its own prediction.

### Reinforcement Learning

Reinforcement Learning uses a weaker form of supervision: the learner (agent) is given neither goal nor directions by a human supervisor. The RL-agent must be programmed to interact with its *environment*, which in turn must:

- inform the RL-agent about its current state,
- yield a scalar reward for the last action.

This is repeated many times [RL-agent performs action, receives state and reward information from the environment]. During many iterations a RL-agent typically tries to maximise future rewards by adjusting its action policy based on the (cumulative) information from the environment.

Both areas are relevant for *Listen and Babble*;

1. Supervised learning is used to train an auditory system to correctly classify speech sounds, while
2. Reinforcement Learning is used by a computer simulated infant learner to acquire the ability to reproduce (or imitate) such speech sounds.

In the next paragraph, I attempt to explain the full model since my research also uses the same model.

In some of the following sections, while summarising how *Listen and Babble* works, its achievements and prevailing difficulties, I will, on occasion, be using words and concepts that are only introduced in section 2 (Methods). In such cases I ask my reader to hang on and 'live with the gaps', or skip to the relevant sections before returning again to these.

#### 1.2.2 The model

I also ask the reader to recall that we looked at two challenges (or stages) in infant speech acquisition, namely: perception and imitation. The *Listen and Babble* model works on the assumption that speech perception (right hearing of speech sounds) is acquired *before* the infant actually imitates. This would mean that we can treat these two separately. Murakami, in his model *Listen and Babble*, first trained an *Auditory System* to accurately categorise speech sounds as one of 4 classes:

*/a/, /i/, /u/, null*

The vowel classes signify German vowels and 'null' class simply stands for neither */a/, /i/ or /u/*. Computer-synthesised speech samples were produced

with VocalTractLab (explained in section 2.1), then individually listened to and labeled by hand with either */a/*, */i/*, */u/* or *null*. The auditory system was trained using *Supervised Learning*. Remember that in Supervised Learning, the learner is trained using many correctly labeled sets of data and learns to make its own predictions for new, unlabeled sets of data<sup>8</sup>. In this case, the dataset consists of sound files (labeled by the human user), and the auditory system is trained to accurately classify speech sounds.

Training the Auditory System models the perceptual stage, where infants learn from ambient speech to recognise language-specific speech sounds and learn to form the same linguistic boundaries (e.g. between vowels) as their caregivers would have. The second stage (imitation) is done using *Reinforcement Learning*.

Recall that infants seem to explore their own speech-instrument (vocal tract), trying out all kinds of gestures and reinforcing such which yield 'promising' speech sounds – sounds which sound 'right' in the infant's linguistic setting (parents, siblings, local communities, etc). The Reinforcement Learning agent (RL-agent) forms a vocal gesture, uses a synthesiser to form an output (speech sound) which then in turn is analysed by the (already trained) auditory system. The auditory system then returns a reward signal to the RL-agent: high rewards if the speech sound was recognised as either */a/*, */e/*, or */u/*, low rewards if the sound wasn't recognised (class *null*).

Imagine a human infant babbling (forming its own vocal gesture), then hearing its own speech sound. If the sound was recognised as similar to the caregiver's speech, the infant will be delighted and continue to form similar vocal gestures.

The RL-agent interacts with (passes on articulator positions to) a simulated vocal tract. The vocal tract creates speech sounds using those articulator positions. The speech sounds are then perceived by the auditory system (which has already been trained to perceive correctly!). The auditory system passes a confidence vote, something like: "How much did this sound really sound like an */a/*?" A reward is computed from the confidence vote and the RL-agent updates its policy based on the reward and executes a new vocal gesture by sending new parameters to the vocal tract. This is repeated ad infinitum, or until rewards become satisfactory (we have learned the given target vowel).

The model's two stages (perception and imitation) are illustrated in Figure 6. The synthesiser (vocal tract model), the auditory system and the RL agent / algorithm are discussed in more detail in sections 2.1, 2.2 and 2.3.

---

<sup>8</sup>Take, for example, the task of predicting the share value of a certain firm. Here, the state is simply the value of the share. If we mined numbers of certain twitter keywords mentioned in connection with that firm, we might use this 'twitter-mood', together with the current share value, to predict the next value of the share. In this case, twitter mood + current share value for a certain timepoint is one dataset. The following share value (say, 1 day afterward) is the state. Take many such datasets + states (e.g. from the last 3 years) and train a computer to make accurate predictions for future states.

These predictions should be reasonably reliable, and the said firm has a good early-warning system (assuming a stable market!).

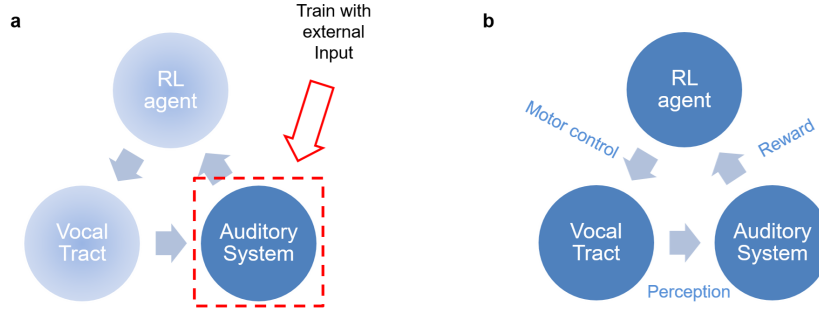


Figure 6: **a:** Perceptual stage of *Listen and Babble*. The auditory system is trained with external input (hand-labeled speech samples).

**b:** The imitation stage of *Listen and Babble*. The RL-agent passes a set of motor parameters to the Vocal Tract (the VTL synthesiser) which in turn produces the resulting speech sound. This sound is listened to by the (already trained!) auditory system and a reward is passed back to the learner for each set of motor parameters. This process continues until a reward boundary is achieved for a targeted vowel.

### 1.2.3 Some results

Murakami managed to successfully train the Auditory System to classify  $/a/$ ,  $/i/$ ,  $/u/$  and *null* with an accuracy of approx. 94%. Part of the Auditory System involved in training is the *Echo State Network* (an Artificial Neural Network). For ESN reservoir sizes (number of artificial neurons) of over 100, the error rate plateaus. (An untrained ESN would, on average, have a random classification accuracy of around 24%). In Figure 7 a, the confusion-matrix is displayed for an Echo State Network of 1000 neurons.

In a second step a RL-agent learned to imitate vowels  $/a/$ ,  $/i/$  and  $/u/$ . Remember the complexity of finding the right position for each articulator (tongue, jaw, etc.) in order to produce a correct phoneme? In *Listen and Babble*, a 16-dimensional motor space is explored<sup>9</sup> in order to find vowel-producing positions of the articulators (more about the articulator parameters in section 2.1).

In the case of  $/a/$  and  $/i/$ , all 16 motor parameters (articulator positions) were learned, each articulator starting from an initial, neutral position (the position of the articulators when producing German phoneme  $/@/$  as in *'viele'*).

<sup>9</sup>This is not uncommon for Reinforcement Learning problems, which frequently deal with huge numbers of possible actions (or *large action-spaces*).

On occasion, a reinforcement learning problem will enable a RL-agent to visit all possible states multiple times, yield a (sometimes delayed) reward for each of these states, and compute a 'value' of that state, or state-action pair ("how much reward we might in future expect, when we're in that state and perform this action"). But often, the RL-agent will only be able to visit a few of these states, simply because there are so many. Take, for example a RL-agent learning to play chess. There are approx.  $10^{55}$  possible ways to place the chess pieces on the board! In that case, a Reinforcement Learning algorithm would have to cope with the fact that they will only be able to explore a very small number of possible states. This is also so in the state space of a human vocal tract.

One important finding was that the model couldn't learn  $/u/$  with all 16 degrees of freedom. The learner needed to keep jaw and lip parameters (3 in total) preset with the 'mentor parameters' (the known parameters from a standard VocalTractLab speaker). The reason is that  $/u/$  requires quite extreme lip and jaw positions, which are simply not found by the RL-agent. The fact, however, that the learner needs to know lip and jaw positions (and not derive them from exploration only), is not too far-fetched. The lips' shape is a conspicuous visual feature in the face of a caregiver, which could easily be picked up by the infant and associated with the simultaneous speech sounds. Thus, visually guided learning (13 motor parameters) proved successful when learning all three vowels, including  $/u/$ .

Figure 7 provides us with a visualisation of the correct-classification rates of the auditory system (a) and the quality of the learned vowels compared to mentor vowels in formant space (b) for learning with either all parameters (16) or visually guided learning (13 parameters).

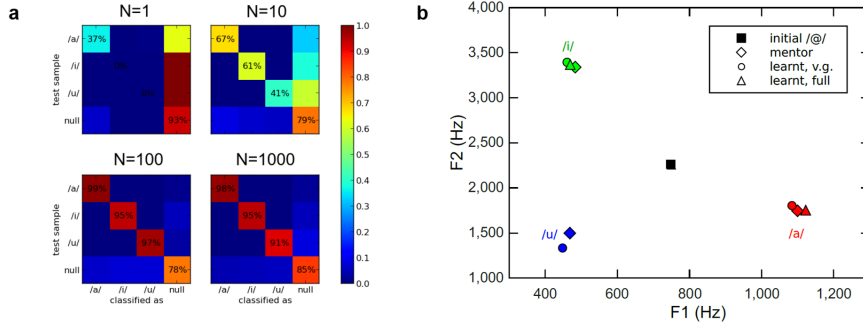


Figure 7: **a**: Confusion matrix of an ESN of  $N_{\text{neurons}} = \{1, 10, 100, 1000\}$ . A confusion matrix shows what fraction of testing samples of each class are classified as that class. Few vowels were classified as null class (or other vowels), however 15% of null-speech sounds were (wrongly) classified as one of the vowels. An ESN manages to correctly classify all 4 classes  $/a/$ ,  $/i/$ ,  $/u/$ ,  $null$  with an accuracy of approx. 94% when using more than 100 neurons.

**b**: Imitation learning in formant space. The agent starts with 'neutral' articulator positions yielding  $/@/$  sounds (*square*) and then moves on to imitate the mentor vowels (*diamond*), producing acoustically similar vowel versions when learning with all parameters (*triangles*, – only  $/i/$  and  $/a/$  learned) or visually guided 13 parameter learning (*circles*).

#### 1.2.4 Open questions

In this section, I will try to show a selection of questions that were still left open in the original *Listen and Babble* model. I specifically mention these in order to show how my work takes up these questions, rules out certain approaches and offers further suggestions.

**Speaker Normalisation** Just like other models dealing with speech perception, *Listen and Babble* faces the problem of *speaker normalisation*. When the

infant’s auditory system was only trained on adult speech samples, its own speech was not recognised (we discussed this challenge of speech perception in section 1.1.1). So that the model would work, the auditory system was trained on both infant and adult speech. But obviously an infant would not have the opportunity to perceptually train on correctly pronounced vowels from his own vocal tract!

We could however argue that the *Listen and Babble* model should be applied to speech acquisition in a social context. We might interpret the auditory system not as that of the infant itself, but of the (fully trained, and already able to generalise across speakers) caregiver auditory system. The rewards would then no longer be interpreted as only inner nerve and hormone signals but as social feedback (approving sounds, – rewards given by the caregiver).

In my opinion, however, this interpretation forfeits some of the advantages of this model over other models of speech acquisition (most importantly the fact that *Listen and Babble* offers an explanation as to why infants also take to babbling without social stimulus, or direct caregiver influence [30]). It makes sense to think of the auditory system as being the infant’s, giving rewards for speech-like sounds without needing the ever-present interaction with a caregiver.

In this thesis, I sought to make some first steps in two directions:

- How well does speaker generalisation work in connection with echo state networks (part of the auditory system)?
- How will learning be affected by an auditory system trained to generalise over a range of ages and both genders?
- Should we change our approach (two-stages, target acquisition before imitation) in order to tackle the problem of learning speech generalisation?

**Small steps toward language** *Listen and Babble* only learns 3 very distinct vowels. Those vowels, /a/, /i/, /u/ are the most distinct German vowels, so it seemed natural to start with these and it obviously is just the first of many steps toward actually perceiving and imitating language. One might ask: What if vowels that are phonetically more similar were introduced? Will the auditory system still be able to distinguish well? What effect would eventual lower sensory accuracy have on learning?

In this thesis, I introduced two further vowels (/o/ and /e/, which are similar to /u/ and /i/ respectively) and sought to make the project implementation more flexible for even implementing arbitrary speech gestures<sup>10</sup>.

---

<sup>10</sup>See Readme.md in the *Listen and Babble* repository, “Towards arbitrary vocal gestures” [24].



## 2 Methods

In this section, I describe the tools used in the model *Listen and Babble*, focusing on changes I made in my research. For a more detailed description of the reinforcement learner, I refer to Murakami’s thesis.

### 2.1 VocalTractLab

*Vocaltractlab* (VTL) is a state-of-the-art speech synthesiser [11]. VTL produces speech sounds based on a 3-dimensional vocal tract model. The model can be controlled using 20 articulator coordinates. Once a certain articulator configuration is set by the user, the airflow through the open space (which is given by the position of the articulators) is simulated. This airflow consists of pressure waves produced by a model of vocal fold motion. Solving the corresponding differential equations of air passing the hyoid, velum, being guided by the tongue and lips results in a life-like (though artificial) speech sound. A 2-dimensional cut through VTL’s vocal tract model is shown in Figure 8b. The software comes with a GUI (Graphical User Interface), where the user can hand-position the articulators and directly produce the corresponding speech gesture.

Most speech synthesisers produce sound based on speech recordings. VTL is different in that the (somewhat simplified) process of speech production in the human vocal tract is simulated. This is convenient for our cause, since we can enter a set of articulator parameters (or *motor parameters*) and VTL produces a speech sound based on those parameters.

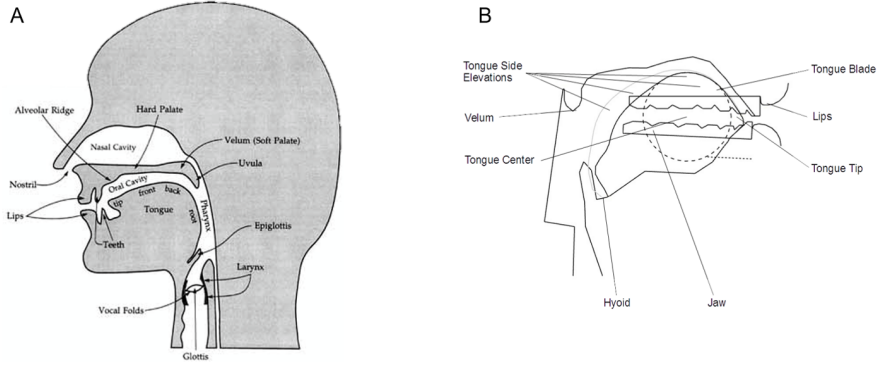


Figure 8: **A:** Articulators of the human vocal tract. Taken from [32]. **B:** 2-dimensional cut through VTL’s vocal tract model.

The Reinforcement Learner uses VTL as a part of its environment. 2 of the 20 parameters controlling the articulators are constant for German vowels (velic opening and horizontal jaw position) and thus will not be learned. Two more parameters (tongue root coordinates) also don’t count as learning parameters since they are derived from other motor parameters. The remaining 16 parameters can be individually explored by the RL-agent by simply giving a set of parameters to VTL, which synthesises a speech sound that may or may not

sound like a German vowel. For such applications, VTL comes with a Linux-API callable in Python and Matlab where VTL can be called directly in any given Python/Matlab code.

The remaining 16 degrees of freedom constituting the motor space are:

- 10 tongue parameters: tongue body coordinates (TBX, TBY), tongue tip coordinates (TTX, TTY), tongue center coordinates (TCX, TCY), and tongue side elevations (TS1, TS2, TS3, TS4),
- 2 lip parameters: lip separation distance (LD) and lip protrusion (LP)
- 2 hyoid coordinates (HX, HY)
- jaw opening angle (JA),
- velum shape (VS).

(List taken from [2]).

VTL comes with two preconfigured speakers (one male infant and one male adult speaker). VTL speakers are anatomically constructed using MRI data from human vocal tract sections. One version of VTL (which is still in development) allows the user to interpolate speakers of various ages between infant and adult stage. I made use of this feature when producing ambient speech from multiple speakers of various ages (see section 3.1).

These 16 parameters are limited by the anatomy of the current VTL-speaker. Each parameter will have a different physiological range of possible values (e.g. the jaw will not be able to open more than a certain value which is still considered realistic). We can, however, normalise the parameters to  $[0,1]$ . Motor learning will thus take place in a 16-dimensional cube with edge length 1 (range  $[0,1]$  for each edge).

For both (infant + adult) speakers, vocal tract configurations are available for /@/, /a/, /i/ and /u/ in VTL 2.1. These were fitted to actual MRI data of patients while they were speaking.

We call these sets of parameters mentor or target parameters. I hand-set these target parameters for all non-standard speakers from age 0 to 20. These hand-set vowels (which I call *prototype vowels*) are not based on MRI data, but rather interpolated from vowel gestures of the standard VTL speakers.

A point worth mentioning here, especially in the context of trying to hand-fit speaker anatomies and gestures in VTL, is this: mapping parameter configuration to speech sound is not unambiguous in both directions. Many different articulator positions can produce roughly the same sound. This becomes obvious to anyone who tries out various motor configurations in VTL and notices that there are many different ways of producing, for example, the vowel /i/. In deciding whether a given articulator representation of a speech sound is a good representation, the following question is crucial:

Is the articulator shape realistic (e.g. does the tongue seem naturally positioned)?

In order to ensure this was the case in all my prototype vowels, I checked with the preconfigured vowel configurations of the infant and adult VTL speaker and visually compared the two shapes.

VocalTractLab was used throughout this thesis in two ways:

- Speech samples (ambient speech) were synthesised using VTL. These are labeled and used for training the auditory system of the infant learner. Ambient speech is produced by gaussian sampling near motor parameters of the prototype vowels. Some of these samples will sound like good representations of the prototype (e.g. prototype vowel /a/ of male speaker 4 (age 4)). These are given the label /a/. Some of the samples (especially those farther from the mean), will not really sound like a German vowel and are given the label *null*. The Auditory System is then trained with most of these samples (training set), and tested for accuracy on the rest of the samples.
- VTL is used when babbling (imitation stage). The RL-agent explores a set of (or all of) the 16 motor parameters by choosing values for each parameter. VTL then accepts those parameters and produces a speech sound<sup>11</sup>. In this application, VTL serves as part of the environment for the Reinforcement Learner.

---

<sup>11</sup>The VTL-produced sound might also be silent. This is the case when the articulators are positioned in a way that completely cuts off airflow through the vocal tract.

## 2.2 Auditory System

The auditory system is the part of the model that accounts for perception of sound. Just like an infant has to learn language perception, supervised learning is applied in order to train a neural network to distinguish incoming sounds as different phonetic groups (vowels, in this case). To distinguish (in this context) means: “to what confidence can I classify a sound as belonging to each class?”. After training and when confronted with a good representation of a vowel, the auditory system should return a high confidence for that specific class and low confidences for all the other classes.

Our auditory system works with three components that take us from a sound input to class-confidences:

- The cochlea model simulates sound processing in the inner ear. A speech signal is tonotopically transformed into nerve activations when receiving (tonotopically organised) sound frequencies. We use the dual resonance nonlinear (DRNL) filter model as described in [40]. It is implemented in “BRIAN hears” [41], an extension of the BRIAN neural network simulator for auditory processing [42]. The model covers the range of 100 Hz to 8 kHz for input sound and returns neural activation for 50 channels.
- An *Echo State Network* (ESN) is what lets us simulate auditory memory [43]. A static<sup>12</sup> reservoir (a pool of recurrently connected neurons) connected to the channels of the cochlea model. These neurons correspond to a simplified auditory cortex. Neurons in the reservoir are connected to output neurons (one for each class). The weights of these connections can be learned by presenting correctly labeled samples to the ESN and adjusting the connection weights.
- Classification based on the ESN-readout. The readout is a set of neuron activations over time (bounded by the duration of the speech sound) for each class  $v$  (vowel). After averaging each class’s output activation  $a_v(t)$  over time  $:= a_v$ , a confidence for each class is computed using the softmax function:

$$c_v := \frac{\exp(a_v)}{\sum_i \exp(a_i)}$$

The confidence of one class acts as reward for the reinforcement learner during the imitation phase.

Keep in mind that the auditory system (in particular the ESN) has first to be trained to classify correctly (target acquisition). In the imitation phase we then use the pretrained ESN to return rewards to the reinforcement learner. The RL-agent targets a specific vowel to learn – the confidence of that specific target  $c_{target}$  is then used as reward<sup>13</sup>.

---

<sup>12</sup>Here, ‘static’ means that when learning (after being randomly chosen) the weights in between neurons in the reservoir are kept constant. Usually, neural networks learn all weights between neurons using (e.g.) back propagation. Recurrent neural networks get rid of certain difficulties of learning hidden-to-hidden connections by simply learning the linear weights coupling neurons in the reservoir to the output neurons (class neurons).

<sup>13</sup>To be precise: The reward is computed using the confidence. Other factors like metabolic cost and penalties from overstepping boundaries of the parameter-hypercube are also taken into account. More on this in section 2.3.

Both stages are illustrated, shown with the components of the auditory system in Figure 9. All connection weights that are randomly initialised are done so in a very specific manner. Connections between neurons in the ESN are set in order that input (e.g. speech input) ‘echoes’ around the network for some time. These connections are very sparse (most weights are 0) and highly recurrent (many connected neurons are connected in both directions). This creates lots of loosely connected coupled oscillators. These oscillators (the neurons of which they consist) are coupled to the output class neurons which are learned during the perceptual stage. When imitating, all connections are kept constant (the ESN has already learned to classify).

ESNs can, in principle, use any neuron model. In *Listen and Babble*, non-spiking leaky integrator neurons were used (we can switch from leaky to non-leaky).

One important parameter of an ESN is the size of the reservoir  $N_{res}$  (the total number of reservoir neurons). The classification task is a bit like fitting a (non linear) function to the data (speech sounds). Increasing the number of neurons is similar to increasing the rank of a polinom fit – data is fitted more accurately with increasing rank (or  $N_{res}$ ) but the danger of overfitting also increases.

Another parameter worth mentioning here is the *spectral radius*  $\rho(\mathbf{W})$  (the maximal absolute eigenvalue of the Matrix  $\mathbf{W}$  containing all the reservoir connection weights). As a rule of thumb,  $\rho(\mathbf{W})$  “should be greater in tasks requiring longer memory of the input” [45]. Varying this parameter was not part of my work but could, in future, be part of increasing memory of the auditory system when looking at production of syllables (and no longer only phonemes).

For more on the used ESN parameters, see section 4.2.1 in [2].

In section 2.1 I explained how I produced ambient speech from a group of speakers of different ages. No longer training the ESN on the two standard speakers only but using a wide variety of speech sounds to train for each class lets the auditory system generalise over speech sounds from various ages. The accuracies of such ESNs, trained with many speakers and with 2 more classes (/e/ and /o/) added, are shown in section 3.2.

I only altered the size of the ESN reservoir in perceptual training. All other ESN parameters were carried over from Murakami’s thesis.

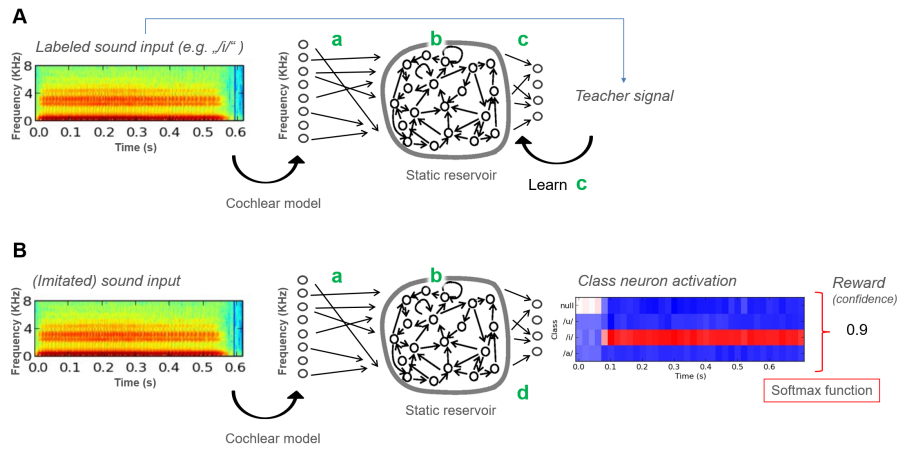


Figure 9: The auditory system in target acquisition (perceptual training) (**A**). During imitation (**B**), the auditory system is used as part of the environment of the RL-agent, returning rewards for speech input produced by the learner. The cochlear model transforms sound into nerve signals of tonotopically ordered neurons (as in the cochlea). These are connected to the static reservoirs using randomly set, constant connections (**a**). Stochastic weights between neurons in the reservoir (**b**) are also kept constant after initialisation. Reservoir-to-output neuron connections (**c**) are learned (linear regression weights of the teacher outputs on the reservoir neuron states) and then kept constant during imitation (**d**).

## 2.3 RL algorithm

An agent learning all the 16 motor parameters mentioned in previous section 2.1 is equivalent to the agent finding a certain point in a 16-dimensional motor space. This needle-in-a-haystack-problem is solved using *Covariance Matrix Adaptation Evolution Strategy* (CMA-ES).

Since my focus shifted towards the perceptual stage in this thesis, I will only very briefly cover the RL-algorithm and point the reader to an abundance of descriptions found in the Web (e.g. [44]).

CMA-ES is a realisation of Reinforcement Learning. An agent interacts with its environment and, based (in this case) only on returned rewards, optimises its behaviour. In our form of imitation we are confronted with a black-box optimisation problem. The learner knows nothing about the physics (or phonetics) happening in his vocal tract. He simply positions the articulators and receives a reward (“how well the produced sound represented a certain target”). CMA-ES is an algorithm formulated for such cases. Points in the parameter hypercube are sampled using a gaussian distribution (the sampled points in each iteration are called *sample generations*). The new  $x_{mean}$  is computed as the weighted average of the  $\mu$  best samples of the last generation. As the algorithm thus finds and converges in local reward-optima, speech sounds of the RL-agent more and more accurately represent the teacher signals (or prototype vowels).

**Target choice** Deciding on which target to learn is an important part of learning efficiently. Humans experience extrinsic motivation (i.e. direct rewards or punishments from the environment) as well as intrinsic motivation (not driven by immediately useful rewards, but rather other concepts such as innate curiosity, or the desire to increase one’s competence). Murakami discusses intrinsic motivation in the framework of Listen and Babble in his thesis [2] (p. 21-22, 30-31).

As in the original model, I arbitrarily set the target at the beginning. However, unlike Murakami, I did not finish learning a target until choosing the next one. The original Listen and Babble model, after achieving satisfactory rewards for a target (i.e. vowel), chose the next target based resampling and choosing the highest confidence target that was not yet learned. Thus, easier targets were learned first until the RL-agent committed himself to targets that are harder to reach. This reflects ideas on intrinsic motivation, where the infant babbler could be motivated by making rapid progress, and leaving slower learning until later, when easier targets have already been mastered.

I made a slight modification in that I let the RL-agent swap targets in each generation (after receiving a reward for the generation of samples and choosing the  $x_{mean}$  for the next generation). If confidences towards other targets are higher, the RL-agent chooses the target with the highest confidence as his next target. “I am trying to imitate /a/. But what I just tried sounded more like /i/. I’ll take that as my target instead!”. This way the agent is more directly driven by intrinsic motivation.

## 3 Results

### 3.1 Ambient speech

#### 3.1.1 Age-specific pitch

Pitch, a prominent feature of speech, changes from infancy to adulthood. I modeled this feature by performing spline interpolation on data from [23, 46]. The actual values used for the speakers are documented in the control parameters script (*control/get\_params.py*) [24] and plotted in Figure 10 a.

#### 3.1.2 Speaker series

Simulating ambient speech for speakers of various ages meant producing many different speaker anatomies. This speaker series was constructed using VTL. Each speaker’s anatomy was calibrated based on a model of anatomical development implemented in VTL. I chose an age difference of 2 years from one speaker to the next. This lets each speaker sound audibly different to the next speaker in the series. Both sexes (22 speakers in total) were constructed using the following ages (in years):

$$\{0(\text{newborn}), 2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}.$$

My speaker series differed from the anatomical model development in VTL in the following points.

- Overall size of the glottis (male/female) was derived from data given in [33].
- Size of the vocal folds (male/female) is based on data given in [34].
- In the six-year-old male speaker, upper molar height was changed from 0 cm (VTL anatomy model) to 0.36 cm in order to give a smoother transition between the 6-year-olds and the 8-year-olds.

For a 3-d visualisation of some of my speakers, see Figure 10 b.

#### 3.1.3 Vowel shape settings

As a second step<sup>14</sup>, vowel gestures were modeled for each speaker. Since each speaker has his own specific anatomy, the set of parameters used for a specific vowel will look different from the previous speaker (or, in general, all other speakers with slightly different ages). Modeling vowel gestures resulted in a

---

<sup>14</sup>These steps are shown in logical progression. In reality, multiple speaker series were created and altered a few times to make them more realistic. For each of these, vowel gesture parameters were set and reset in order to get rid of certain disturbing effects in the audio data (e.g. when airstream was too narrowly restricted, this could produce a scratching noise in the actual data used as ambient speech).



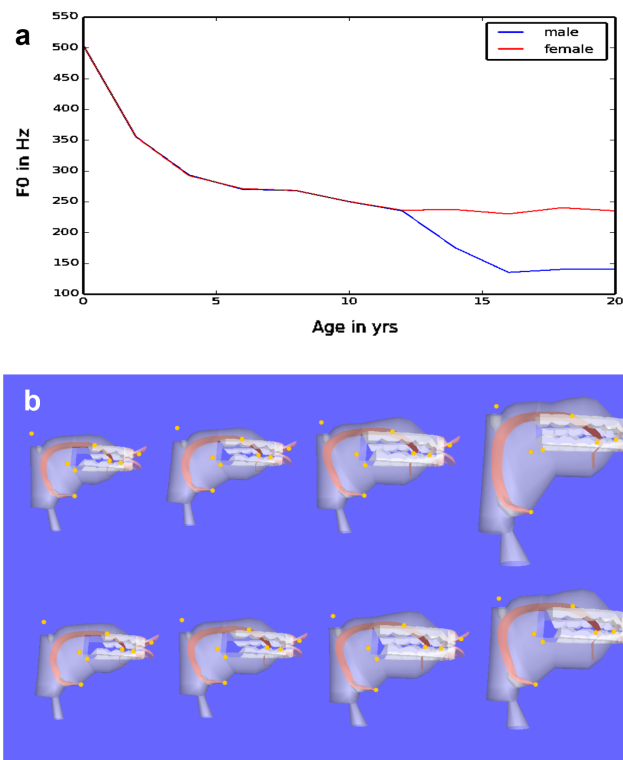


Figure 10: **a**: Pitch values used in the model (both male and female) from 0 to 20 yrs (same age development as in b). **b**: Selected speakers from the speaker series shown in the VTL graphical interface. Top row from left to right: male speakers aged 0, 4, 10, 20. Lower row from left to right: female speakers, same ages.

set of prototype vowels for each speaker. In this thesis, I calibrated prototype articulator positions for the following vowels:

$$/a/, /e/, /i/, /o/, /u/.$$

Vowel shapes were set by hand in VTL. After listening to the resulting (synthesised) speech sound of a hand set shape, I corrected the articulators and again listened to the sound. This was repeated until I perceived the speech sound to be a good representation of a given vowel.

I already mentioned in section 2.1 that often multiple articulator configurations can produce the same speech sound. It is possible to produce all vowels in VTL with unrealistic positions (e.g. of the tongue). A human calibrating a vocal tract in VTL needs to be guided by knowledge of how vowels are indeed produced by humans. I therefore invested some effort to become familiar with the articulators' positions of the standard speakers in VTL for each vowel, since these were set by a trained phoneticist.

I note here that neutral articulator positions ( $/@/$  in SAMPA<sup>15</sup> notation, also called *schwa* as in German 'bitte') are preset in the speaker anatomy model of VTL. We therefore do plot  $/@/$ , but remember that it is not learned by the infant in this thesis, nor is it a distinct class in the auditory training.  $/@/$  positions are actually used as initial articulator position for the RL-agent, from which the RL-agent explores and learns  $/a/, /e/, /i/, /o/$  and  $/u/$ .

**Prototypes** In section 2.1 I described how vowel samples are generated for the perceptual learning. Using articulator positions of the prototype vowels, these are slightly varied in order to produce many prototype-like vowel samples. Sometimes, even the small changes in the articulator positions mean a large change in how I perceived the speech sound. This meant I had to label each sample before training reservoirs (section 2.2) on the data.

In the original model, vowel samples near preset vowel shapes (of adult and infant speakers) were used. The formants of those standard vowels ( $/a/, /i/, /u/$ ) spanned triangles in formant space, the adult speaker's triangle having lower  $F_1$  and  $F_2$  values than the infant speaker's vowel triangle (see Figure 11). Since the first two formants  $F_1$  and  $F_2$  are central in vowel recognition we can actually think of speaker normalisation in terms of mapping age-trajectories through formant space, or "which paths do vowels take in formant space with in- or decreasing age"?

In Figure 11, speech sound formants  $F_1$  and  $F_2$  are shown for:

- VTL-preset adult speaker vowels (black markers)
- VTL-preset infant speaker vowels (blue markers)
- my speaker series' neutral articulator ( $/@/$ ) - sounds.

It is easy to see that the  $/@/$  formants form a trajectory from the infant's  $/@/$  to that of the adult<sup>16</sup>.

<sup>15</sup>*Speech Assessment Methods Phonetic Alphabet* (SAMPA) is based on the *International Phonetic Alphabet* (IPA), the main difference being that SAMPA uses ASCII characters. The schwa sound, is written as  $[@]$  in SAMPA, but  $[ə]$  in IPA notation.

<sup>16</sup>This way of thinking about speaker generalisation assumes static speech sounds. Formant values up to section 3.4 are computed by taking the median formant for that vowel. However,

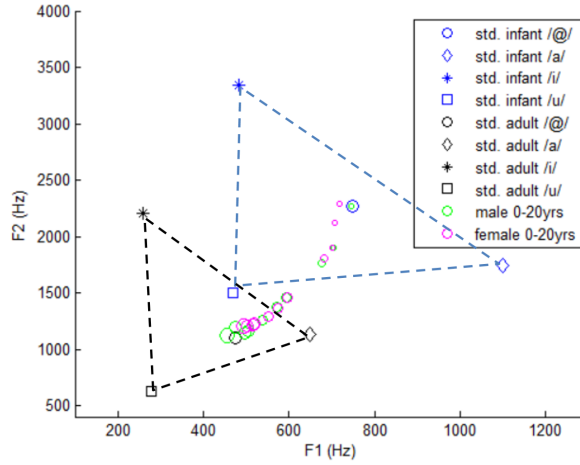


Figure 11: Vowel ( $/a/$ ,  $/i/$ ,  $/u/$ ,  $/@/$ ) formants of the VTL infant speaker (blue), and adult (black). Green and pink circles are neutral articulator positions of my speakers, which form a trajectory from infant  $/@/$  to adult  $/@/$ . Larger circles stand for older speakers. Lines drawn in between vowels serve only to illustrate (vowel triangles).

If vowel prototypes  $/a/$ ,  $/e/$ ,  $/i/$ ,  $/o/$  and  $/u/$  of my speakers also follow similar trajectories as do  $/@/$  sounds in Figure 11, I take them to convincingly represent a development from adult to infant. Thus, when trained on, they ought to work towards normalising vowel perception of the auditory system (bridging the age gap between the infant and adult speaker in perception). An auditory system trained on samples from my speaker series should then be able to correctly perceive vowels from any age group.

Figure 12 a and b show formants of all speakers' vowels (male and female separate) including the neutral position ( $/@/$ ), and Figure 12 c shows all the prototypes (only including vowels used in perceptual training, without  $/@/$ ). Some observations can be made:

- The vowel prototypes show a similar development as the neutral positions do in formant space.
- Some vowel groups strongly overlap (especially  $/o/ \leftrightarrow /u/$ ).
- Vowels from younger speakers (smaller circles) are less regular and further apart.

Despite the fact that shapes weren't set by a trained phonetician, vowel prototypes do resemble vowels recorded by humans of various ages. Formants from most prototype vowels display systematic developments over age, just as is the case in real speech. In Figure 13, a formant space plot is shown for the male

---

while pronouncing one vowel, formants vary over time. This is especially significant in infant-directed speech (with large  $F_0$  variation).

For now, we shall assume constant formants for the duration of a vowel (by taking the median as representative), and take into account changes in formants *over age*.

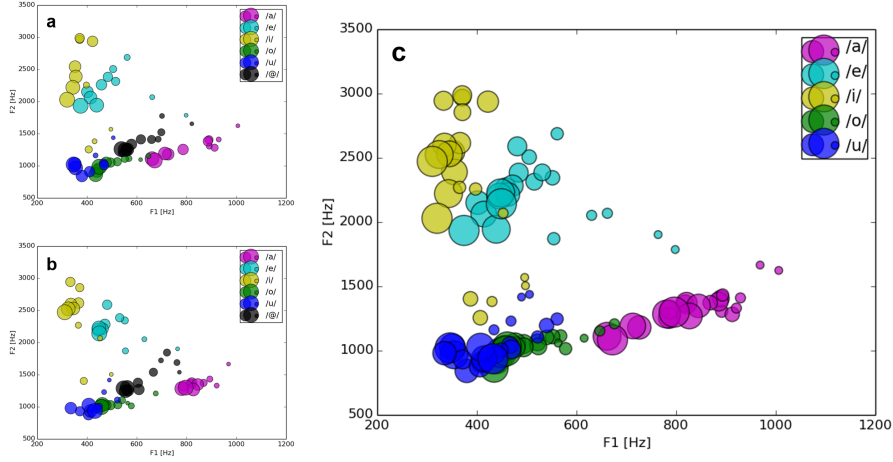


Figure 12: Formants of prototype vowels of my speaker group. Larger circles stand for older speakers. Male speakers (a) and female speakers (b) shown separately, including /@/. c: All prototypes.

and female speakers compared to formants of General American vowels recorded from human speakers of age 5 to 19+<sup>17</sup> [23]. Note that their formants show similarities, although realisations of General American vowels /a/, /eh/, /i/, /u/ are phonetically slightly different to those of German vowels /a/, /e/, /i/, /u/ (shifted in formant space). The following three additional observations can be made:

- Similar vowels (marked with the same colours) take up similar parts of formant space.
- We see a similar development from young speakers' vowels towards older speakers' vowels.
- Though younger speakers again seem less regular as to their formants<sup>18</sup>, this is not seen as clearly in the literature data (with no data for speakers under 5).

### 3.1.4 Vowel Samples

Using the shape parameters (which mark the articulator positions) of the prototype vowels, I sampled near (gaussian sampling with  $\sigma = 0.01$ ) those parameters in order to produce vowel-representative samples, and further away ( $\sigma = 0.2$ ) in order to produce non-representative speech sounds (sounding unlike any of

<sup>17</sup>The group 19+ representing a group of speakers up to the age of 50. Speech development slows down above the age of 20, which justifies such a category. I am not aware of phonetic data of vowels available for children younger than five years (various studies show other phonetic features down to the age of four).

<sup>18</sup>This attribute of the young (age 6 and under) speaker prototypes of my speaker group is actually feasible. Lee et al. also find that younger speakers have considerably more variation in their speech's phonetic features. [23]

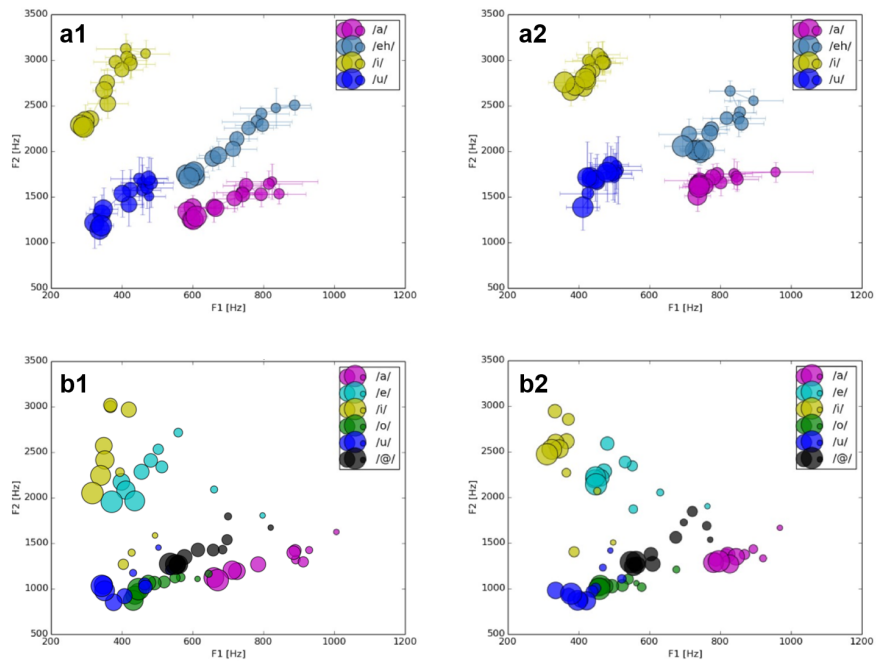


Figure 13: English vowel formants [23] (**a1**: male speakers and **a2**: female speakers) compared to German vowel-prototype formants synthesised in VTL (**b1**: male VTL speakers, **b2**: female VTL speakers). Literature formant error bars are the standard deviations among different subjects' vowels from the same age group. The size of the circle indicates the age (the larger the circles, the older the speakers).

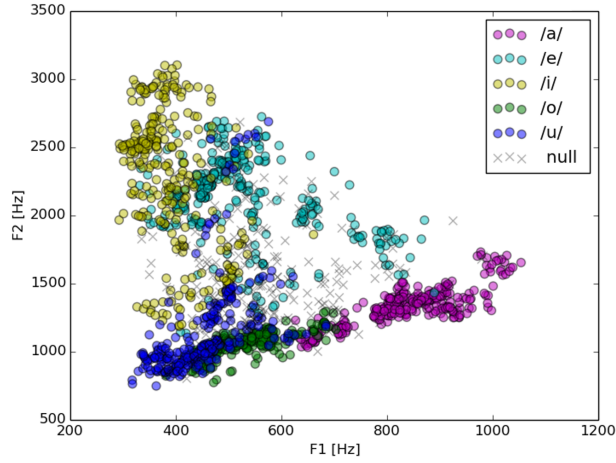


Figure 14: Vowel samples in formant space. *null* represents samples sounding nothing like any of the listed vowels.

the vowels). All these samples were then listened to individually and marked as either */a/*, */e/*, */i/*, */o/*, */u/* or *null*. In order to minimise any biases while labeling, samples were shuffled and then listened to blindly. Between the speech sounds I listened to 1 second of gaussian noise in order to reduce perceptual biases due to the phonetic context.

Although sampled around prototype vowels, samples displayed much larger formant variation. This is seen in Figure 14. We can again make some observations:

- Vowel groups, although sampled near prototype parameters, can be phonetically quite dissimilar (compare with Figure 12 c).
- Sample vowel groups (and thus, the classes which the auditory system must learn to distinguish) strongly overlap in formant space (especially groups */o/* and */u/*, and */e/* and */i/* respectively).
- Of all groups, */a/* samples seem most organised.
- Samples with formants far from the respective 'normal' area in formant space were still recognised as belonging to that group.

## 3.2 Training reservoirs

The produced samples (including their labels) now act as ambient speech for the (modeled) infant. A set of 2112 labeled samples are used in the following steps (16 samples per speaker and per vowel). Auditory systems with various reservoir sizes were trained on the ambient speech. We now train with 6 distinct output classes:  $/a/$ ,  $/e/$ ,  $/i/$ ,  $/o/$ ,  $/u/$  and *null*. 80% of the samples were used to train each auditory system with 20% of samples used as a test set.

The reservoir is trained using the training set and then given the task of classifying the samples in the test set. Rates of wrong classification (error rates) are an easy measure of the accuracy of a specific reservoir.

### 3.2.1 Partial training

What happens if we only partially train a reservoir? In real life, the infant learner should be able to solve the perceptual task without hearing its own speech samples. Solving the problem of speaker generalisation would mean being able to train a reservoir (for the auditory system of the infant) with speech samples other than its own and then still correctly being able to categorise its own speech sounds.

Normally when training a reservoir, test and training set are randomly picked from the entire data set. But what if we remove all samples produced by speakers of ages 0–2 yrs when learning, and then use those samples (0–2 yrs) when testing the reservoir? Figure 15 shows error rates for 20 different reservoir training paradigms, exploring the ability to generalise for specific age groups. Reservoirs with sizes  $N = \{.., 10, 100, 1000\}$  were trained. For each reservoir size (e.g.  $N = 1$ , first row in the plot) errors are plotted (colour) for five different training paradigms. The far left column, for example, shows reservoirs trained with samples from speakers of all ages except 0–2 yrs. Samples from speakers of 0–2 yrs served as test set. Error rates are the mean error rates taken from 30 individual trials. While all subgroups are made up of 4 speakers (2 male and 2 female), the last group (age 16–20) has 6 speakers.

We can observe that:

- Errors are smallest in the center column (omitting speakers aged 8–10). Interpolating thus seems easier than extrapolating (perceiving speech sounds from very young or very old speakers when not being trained on these).
- This extrapolation error (or: inability to normalise) is most strong for young speakers, who display larger phonetic variance in vowel production.
- While on the whole error rates decrease for larger reservoir sizes, this is not the case when extrapolating to young speakers (first column) – the  $N = 100$  reservoirs, trained without ages 0–2 yrs, show smaller error rates than the  $N = 1000$  reservoirs trained without the same subgroup. The larger reservoirs seem to overfit the data in this case.

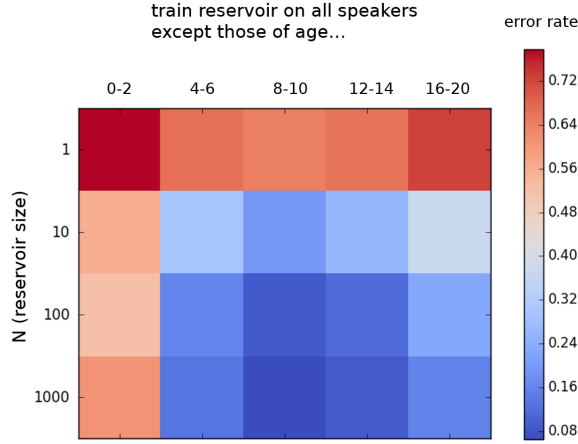


Figure 15: Mean error rates (misclassification rates) over 30 trials for various reservoirs. ESNs of different reservoir size  $N$  were trained using training sets that omit samples from a certain subgroup of speakers. The reservoirs are then tested on that subgroup.

### 3.2.2 Comparing reservoir training paradigms

In order to make meaningful conclusions on how ambient speech from a diverse group of speakers affects the quality of the auditory system classification, I compare four different reservoir training paradigms. Figure 16 shows rates of misclassification plotted over different reservoir sizes. Training paradigms were the following:

1. Reservoirs are trained on only the standard infant and standard adult speaker, discerning four classes:  $/a/$ ,  $/i/$ ,  $/u/$ ,  $null$  (*black*).
2. Reservoirs are trained on the whole speaker group, discerning the same four classes (*blue*).
3. Reservoirs are trained on the whole speaker group, now discerning six classes ( $/e/$  and  $/o/$  included) (*green*).
4. Reservoirs try to extrapolate. Trained on all speakers that are four yrs and older (leaving out 0–2 yrs) but testing on speakers of age 0–2 yrs. These reservoirs attempt a form of speaker normalisation. (Trained on all six classes) (*red*).

Error rates from paradigm one were taken from Murakami’s thesis [2].



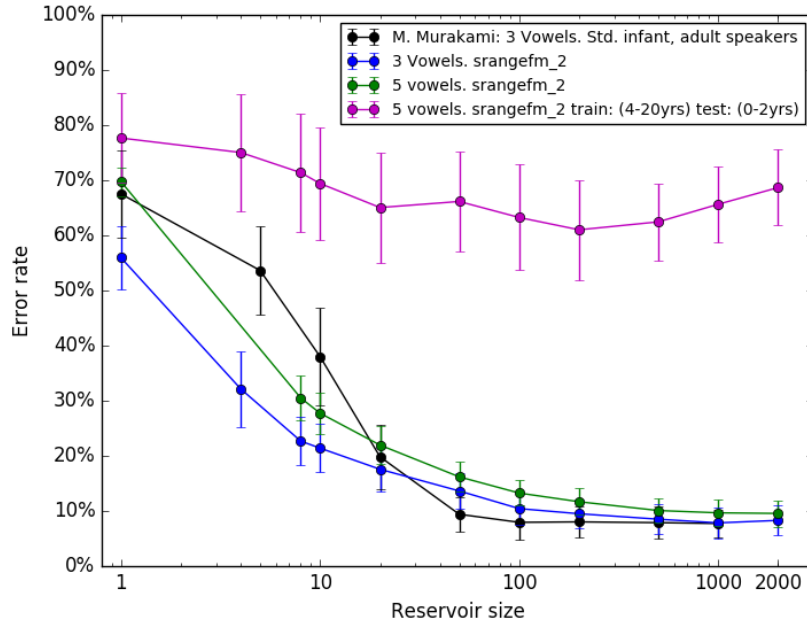


Figure 16: All curves show error rates (rates of misclassification) of the samples' test set for different reservoir sizes. Curve colours code for different training (and testing) configurations.

*Black:* Reservoirs trained on adult and infant speaker samples and four classes ( $/a/$ ,  $/i/$ ,  $/u/$ ,  $null$ ) from [2].

*Blue:* Reservoirs trained on four classes, but using samples from the whole speaker range (0-20yrs, male/female).

*Green:* Reservoirs now trained on six classes (new vowels  $/o/$  and  $/e/$  added), using samples from all speakers.

*Red:* Reservoirs trained on six classes, on samples from all speakers that are 4 yrs and older (leaving out 0-2 yrs) and testing on speakers of age 0-2 yrs. (Extrapolated)

Each data point is the mean error rate of 100 trained ESNs; the error bars are the corresponding standard deviations.

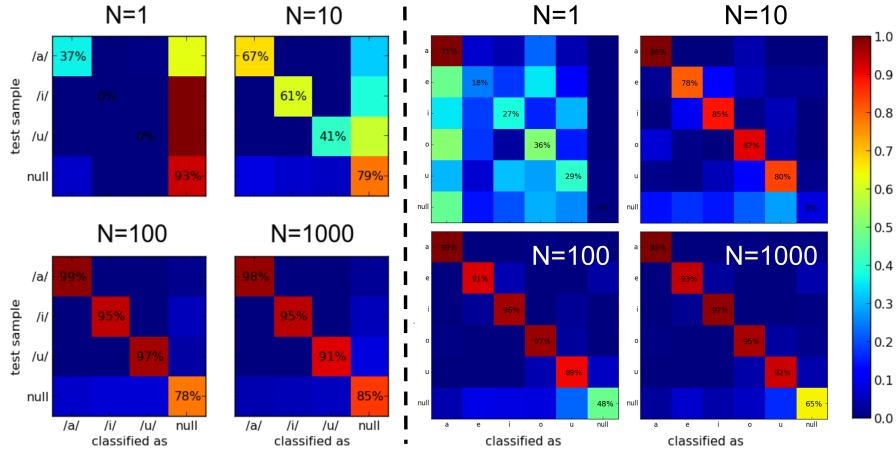


Figure 17: Confusion matrices for Murakami's thesis (*left*) and for this thesis (*right*). When a reservoir is tested for its accuracy, a confusion matrix shows how many samples from the test set (vertical axis) are correctly classified as their original labels (horiz. axis).  $N$  denotes reservoir size.

Observations:

- The increased speech variety of a speaker group (compared with only two speakers) actually makes correct classification easier for smaller reservoirs. However, reservoirs of  $N > 50$  start to efficiently fit data from only two speakers. Differences vanish for larger reservoirs  $N > 500$  (compare *black* with *blue* curve).
- When discerning two more classes we do not see a significant difference in error rate (compare *blue* with *green* curve). We must note that the chance-classification error rate would also be slightly higher for six classes (83%) than for only four classes (75%). Thus, the error rates per class would not be significantly different when comparing paradigm 2 and 3.
- Reservoirs have difficulties classifying correctly when extrapolating towards young speakers (*red (magenta)* curve). Very large reservoirs seem to overfit the data.

In Figure 17, I compare reservoir training paradigm 1 and 3. Despite having to classify more vowels and from a larger variety of speakers, vowels are classified with a correct classification rate that is comparable to Murakami's reservoirs. The main difference is the *null* class: Samples which were labeled as *null* by me were frequently (in 35% of all cases!) classified as one of the vowels by the reservoirs. The reservoirs are “too tolerant” about speech sounds; bad quality vowels are given higher confidence levels than they should be given. Also, vowel groups that strongly overlap in formant space (see, for example, Figure 12) are more frequently mixed up (groups  $/e/$ ,  $/i/$  and  $/o/$ ,  $/u/$ ).

### 3.3 Imitation

Due to time limitations, I do not have statistically relevant results as to how well the imitating RL-agent performs on a generalised auditory system (as in section 3.2.2, paradigm 3). However, I will show the result of one simulation using a reservoir of size  $N = 100$ , trained on all speech samples from my speaker group.

The RL-agent was the VTL infant speaker, whose speech samples are not included in the reservoir training. The agent learned 13 motor parameters, using mentor lip positions and jaw opening parameters. This corresponds to the case of visually guided learning (see [2], section 2.2.3).

I did not use the FIAS cluster and did not yet implement the algorithm in a parallel fashion<sup>19</sup>. Results are shown for 10,000 samples (1000 generations of 10 samples). The RL-agent managed to approximate the target vowels so that they were easily (though not perfectly) recognisable as that vowel when I listened to them. Results are shown for the parameters yielding the best rewards from among all the samples (not necessarily the configuration at the last generation!). In Figure 18, learned vowel spectrograms are compared to prototype vowels. In Figure 19, formants  $F_1$  and  $F_2$  are plotted for prototype vowels and learned vowels. Unlike in Murakami’s thesis, the prototype vowels do not act as direct targets or mentor configurations. Murakami used samples near the articulatory configurations of mentor vowels of the infant speaker himself (along with those of the adult speaker) for the perceptual training. I did not use samples from the (RL-agent) infant speaker. The auditory system is trained on samples from other speakers (Figure 14 showed their samples’ distribution in formant space).

We can observe that learned vowels span a smaller vowel space than hand-set vowel prototypes. The babbler seems to be happy with approximating the vowels. With the exception of  $/i/$ , vowels seem to be nearer to the neutral position and thus not articulated as distinctly as prototype vowels are.

Since CMA-ES learning is a stochastic process, this result should be reproduced and statistically reliable data gathered (e.g. mean learning time for each vowel to yield a certain confidence level). I sometimes even achieved successful imitation for less than 1,000 samples per target (roughly 4,000 samples for all 5 targets).

Another observation is that learning seemed to work better when learning more targets. Remember that the RL-agent is intrinsically motivated, switching from one target to the next. This could mean that learning  $/o/$  could aid in learning (phonetically similar)  $/u/$ . This too should be subject to further investigation.

---

<sup>19</sup>This is easily done. A proposed implementation of the RL-learning is noted in the documentation and in the source code [24].

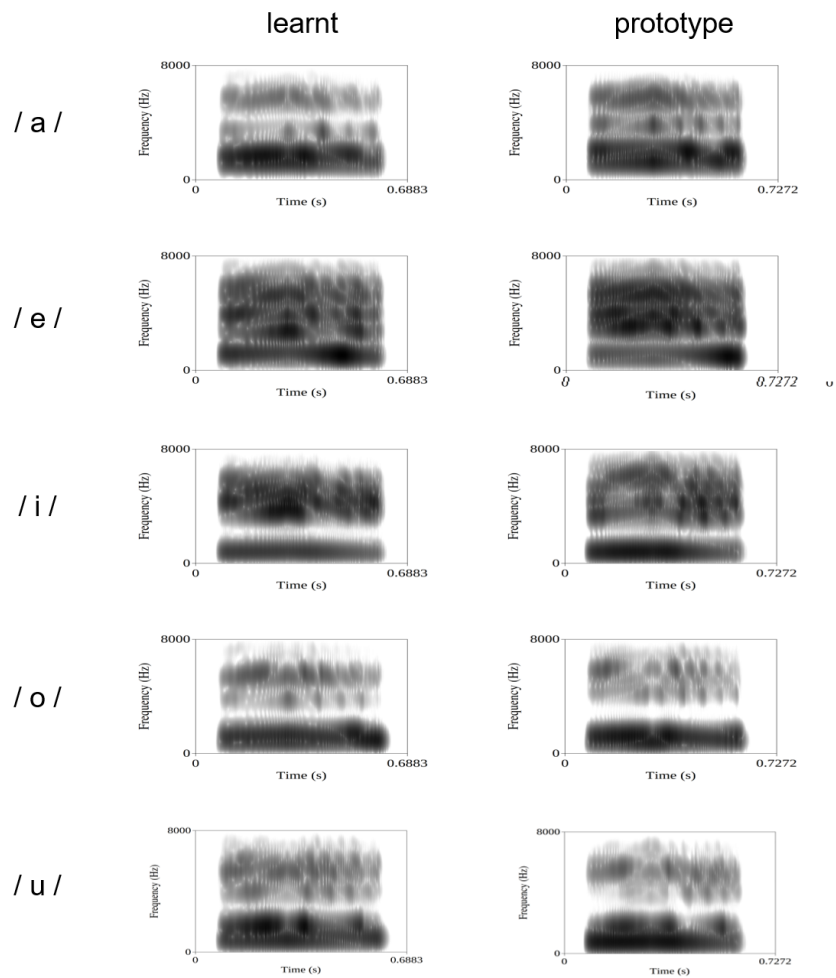


Figure 18: Spectrograms of learned vowels (maximal reward in 1000 iterations of 10 samples) compared with prototypical vowels from the VTL infant speaker as learner.

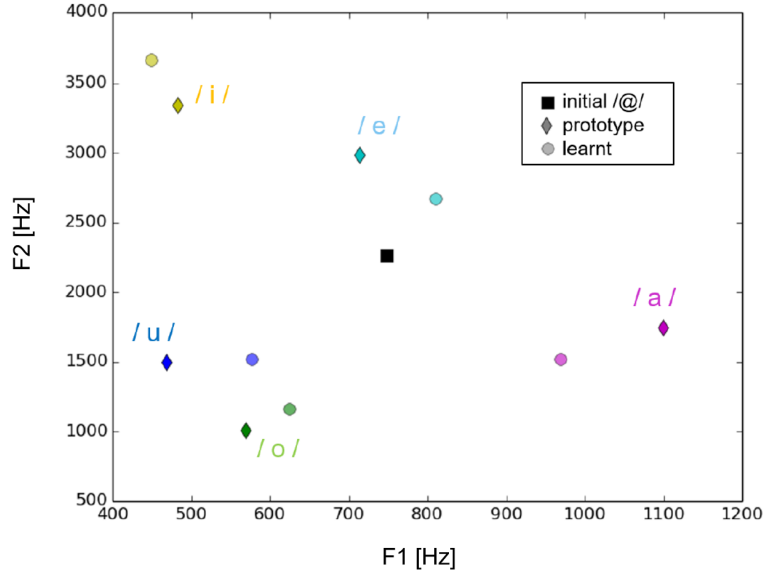


Figure 19: The learner no longer has one perceptual target, but is trained on speech from all different ages. Learned vowels in formant space (weakly) resemble prototypical vowels (preset in VTL). Learned vowel space is smaller than prototypical vowel space (nearly all learned vowels are closer together).

### 3.4 Motherese

As we saw in 1.1.3, motherese helps infants solve the perceptual problems in language learning. Infant directed speech could be integrated in *Listen and Babble* and might even help solving speaker normalisation.

I tried to include motherese-like pitch variation in VTL samples. But labeling VTL-synthesised speech with pitch variation over one or two semitones proves difficult: what sounds like a specific vowel changes and seems to better represent another when pitch (alone) varies too much. This perceptual experience is confirmed by Miller [37] and by Slawson [38] in their research. Miller, for instance, found vowel category boundary shifts in formant space for most English vowels when doubling the pitch (one octave).  $F_1$  boundary can shift from 100 Hz to 200 Hz for a similar change in pitch (200 Hz) [39].

I noticed this perceptual shift, especially when hand-fitting vowel shapes in VTL for very young speakers ( $F_0$  between 400 Hz and 500 Hz). I included slight pitch variations in all samples (two semitones), but noticed that vowels seemed to shift in the way I perceived them (most often between /i/ and /e/ and between /o/ and /u/ or /@/).

When using high pitch modulation in speech we noticeably constrict our own vocal tract, especially in the area of the pharynx. We never only change the pitch, but “accompany” pitch change with the whole vocal tract in order to perceptually stay in the same vowel category. This means that we would have to modulate not only pitch but also some (if not all) articulators in VTL. Laying out trajectories for articulator parameters over time in order to stay within the

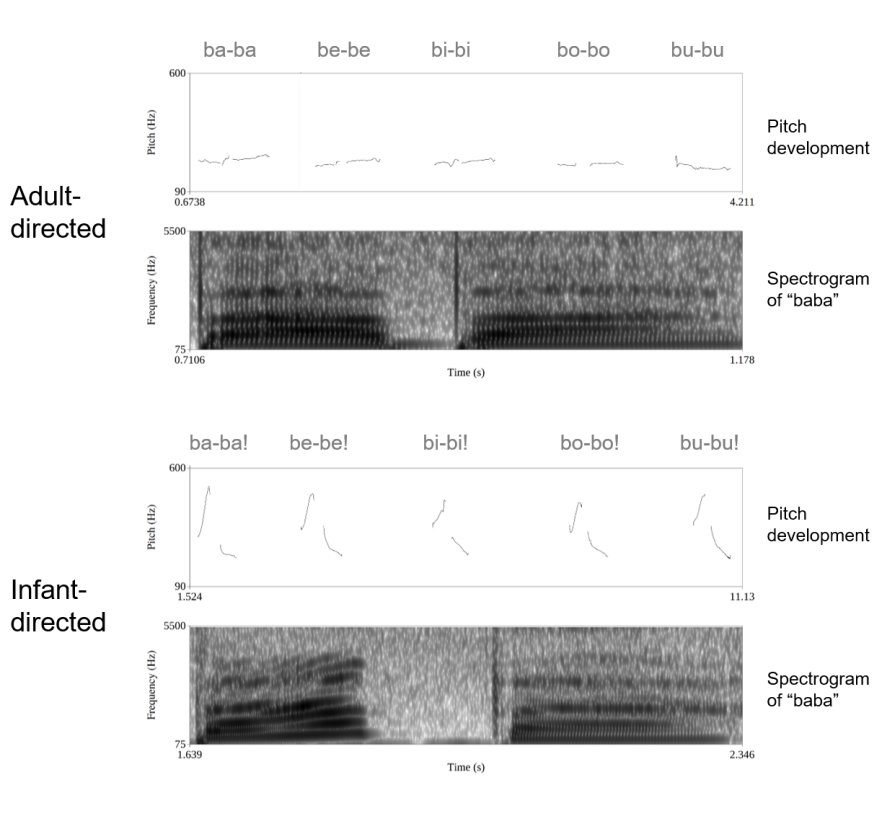


Figure 20: Pitch ( $F_0$ ) variation in adult-directed speech and in infant-directed (motherese) speech. Spectrograms shown for the first “word” in the sentence.

same vowel category the entire duration of the speech gesture is not easily done.

Instead, I pursued the possibility of using human motherese speech samples to train the auditory system. My wife kindly provided me with speech samples of German vowels built into the following linguistic context:

“*Baba – bebe – bibi – bobo – bubu!*”

The sentence was first spoken in an adult directed way, then in an infant-directed way. Figure 20 shows  $F_0$  of the entire sentence and the spectrogram of the first “baba” in adult- and infant-directed speech. It is obvious that not only  $F_0$  but also the formants show large alteration throughout the speech gesture in motherese speech.

I found that extracting formants is not trivial in a linguistic context<sup>20</sup>. My own, male speech proved difficult to read formants from but I managed to produce the trajectories for the female-spoken vowels. The result was larger and more distinct formant trajectories in formant space for infant directed speech, see Figure 21.

<sup>20</sup>Visible horizontal lines in the spectrogram are not necessarily the actual precise formants!

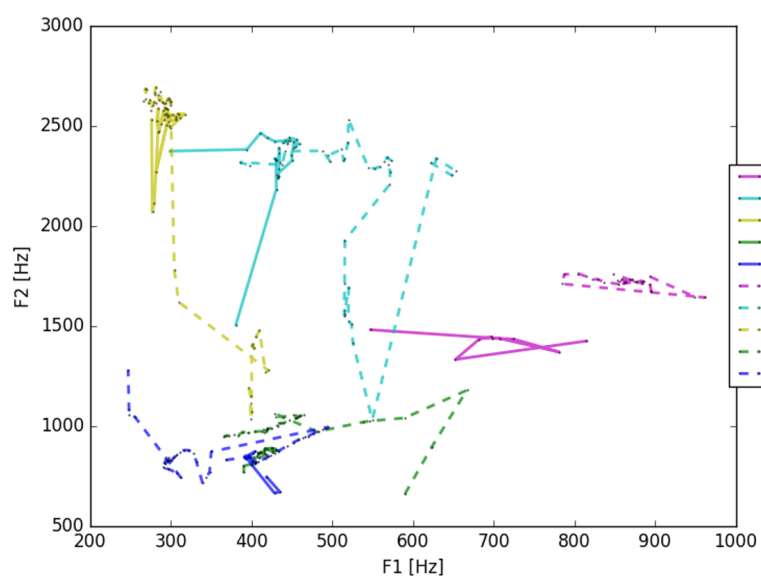


Figure 21: German vowels spoken infant-directed (ID) and adult-directed (AD) in phonetic context. ID speech is marked by larger formant space trajectories and larger vowel space. AD vowel /o/ is hardly visible, and shows no formant development at all. Also, in this case of adult directed speech /u/ and /o/ prove almost impossible to discern using only the formants. Formants of /o/ (AD) are located very near those of /u/ (AD) and are not so clearly seen because they are nearly static.

## 4 Discussion

In this section, I discuss what I think my work demonstrates, and also in which direction I believe my results point, regarding future work.

**Implementation of *Listen and Babble*** The model went through a large amount of changes concerning implementation (code). The project structure should enable more straightforward ways of introducing new steps (like those I will cover in this section). For a more detailed understanding of how the code works, I refer to the documentation on the *github repository* [24].

**Speaker normalisation** The thrust of my research shows: although using a range of speakers increases reservoir classification accuracies of smaller reservoirs, the model still cannot work without perceptually training on its own speech. Trying to extrapolate towards younger speakers with reservoirs trained on older speakers proves more difficult than vice versa, as we saw in section 3.2.2. Speaker normalisation is still a problem for this model of speech acquisition.

**Phonetic accuracy** As mentioned in section 3.1.3, vowels were labeled simply by me listening to them. Since this is subjective, a next step would be to have them labeled by a phonetically trained expert. However, I do not think that this will influence the learning very much, except if the labeling were done more strictly (rejecting more obscure speech sounds as belonging to the null class).

**Sampling strictness** Samples were hand-labeled after being listened to. Like this, even speech sounds that weren't quite perfect, but fell into a specific perceptual category (e.g. were heard as /a/) were admitted as samples for the reservoir training. This, one could argue, is problematic for two reasons:

1. Samples in formant space are no longer distinguishable and categories overlap strongly. This may confuse the classifying reservoir, especially when trying to extrapolate to younger speakers.
2. Infant directed speech in reality is pronounced deliberately, using even more distinct pronunciation (for example larger vowel space). *Listen and Babble*, in a way, tries to form perceptual targets while listening to (at best) adult directed speech.

It would certainly be admissible to only train with samples that show very clear phonetic recognisability. One might, for example, remove all samples (or classify them as belonging to the *null* class) that have formants that differ too much from their corresponding prototype formants. Thus, samples would become better separated in formant space and phonetically more easily distinguishable.

A question might be: when samples are more distinct from each other, does the auditory system interpolate better? In order to test this, one could repeat (with the new, 'stricter' samples) reservoir training paradigm number 4 (from section 3.2.2) and compare error rates with those in Figure 16.



**Motherese** Larger, and more distinct formant-*trajectories* do seem to make infant directed vowels easier to distinguish from each other (especially /o/ and /u/ as seen in Figure 21).

By only looking at the one recorded motherese sample in section 3.4 we can not make significant claims as to whether motherese acoustic features move nearer to those of the infant, and thus help the infant learner to generalise. However, we can see that trajectories are more distinct. Larger (and thus clearer) trajectories in formant space could aid in the generalisation by adding another characteristic, by which an infant (or an ESN) can distinguish a specific vowel better, independent of the precise location in formant space. In the sample, we saw infant-directed /u/ and /o/ mainly distinguishable by their trajectory, and not by their position in formant space.

Using vocal gestures with large pitch variation over time in *Listen and Babble* (while still using VocalTractLab for training data) does not seem feasible. We should consider using data from human mother-infant speech. KIDS (Konstanz prosodically annotated infant-directed speech corpus) [47] seems like an ideal source, since it consists of German speech (spoken to infants by their mothers). Individual syllables are already labeled and one only has to extract those phonemes from which perceptual targets are formed (e.g. only vowels).

**Understanding imitation more fully** For my thesis, I (frustratingly) lacked the time to fully investigate the second stage of the model (imitation). Murakami showed that, when the auditory system only trains on vowel samples produced by the very same vocal tract as that of the imitating agent, that agent is able to completely learn those vowels and attain a near-perfect target match. It is now important to understand how learning works when no longer one perceptual target is formed, but rather a more nebulous perceptual target over multiple speakers. I had time to run some first tests using a reservoir that was trained with all speakers. Further questions are:

- How does the RL algorithm compare with random babbling (sampling random points in the hypercube until a point is found that yields a high enough reward)?
- On average, how many samples do we need for each vowel in order to reach a certain reward level? (Maybe learn multiple targets in parallel instead of intrinsic motivation for better comparison?)

Implementing variable reward thresholds might be a way to more evenly learn targets. We might, for instance, learn all targets up to a low threshold (e.g. 0.2 – 0.3) and only then try to improve utterance by raising target reward thresholds individually.

Also, the current target could be chosen on the go based on the increase in reward (learning progress of that specific target). The RL-agent could train where reward increases fastest, thus focusing on making easier progress first.

**Interdependence of articulator positions** A point mentioned by Murakami in his thesis is the possibility to introduce (biologically realistic) constraints as to what positions each articulator can take. For now, the RL-agent can freely

choose each parameter, independently to the next, resulting (for instance) in extreme tongue shapes. Making articulator parameters interdependent could make learning simpler by reducing the dimensionality of the problem.

I suggest restricting the curvature of the tongue (in the transverse plane) to smaller values (e.g. each only moving between 0.4 and 0.6 in relative coordinates, 0.5 being the neutral flat position) as a first easy step.

**Comparing with other models** The *Listen and Babble* model of vowel acquisition in infants comes with obvious advantages over some other models out there.

Murakami describes some of these advantages in [1]. Here, I'd like to compare our approach in *Listen and Babble* with two models not yet mentioned in Murakami's discussion in his thesis. Murakami provides an interesting comparison with the following models: DIVA, Kröger, Warlaumont, Frier-Oudeyer (see pages 56-57 in [2]).

The reason I mention the following models is this: they both emphasise the role of the caregiver, which in *Listen and Babble* is limited to the forming of ambient speech for the perceptual training. In our model the caregiver is not needed when imitating. This will be the biggest difference between *Listen and Babble* and the following models:

Asada's group [48, 49], due to their research field (robotics), recognise the *correspondence problem*. For instance, in phonetics, they modeled vowel acquisition using an artificial articulatory system (with five degrees of freedom). Because this articulatory system is structurally very different to that of a human, their speech robot faces difficulties finding actions that correspond to human actions with similar results. Training a simple speech robot to match a caregiver's phonetic categories, they show that the robot solves the correspondence in this way: caregivers' speech is not imitated directly (as in "how can I reproduce the sound which my caregiver made?"). Rather, the caregiver imitates the robot's cooing, which in turn enables the robot to learn more vowel-like articulations by mapping the phonetic regions it can produce to those that the caregiver uses to imitate. Through unconscious anchoring in the caregiver's imitation the robot bridges the correspondence problem. The robot's phonetics approached those of the caregiver without actually imitating the caregiver's speech.

The I.S. Howard and P. Messum model ELIJA [51, 50, 52] expands Asada's approach from vowel acquisition to learning syllables and first words. They stress the fact that infants experience the same correspondence problem (between their infantile vocal tracts and those of the caregivers). Their model ELIJA learns equivalence relations between its own vocal actions and the caregiver's speech in response to those actions. ELIJA does not listen to the sound it produced as our model does.

In my opinion, Howard and Messum make strong arguments for the importance of caregiver imitation<sup>21</sup> in solving the correspondence problem (which is similar to solving the problem of speaker normalisation). But I do not know of ELIJA accounting for spontaneous infant babbling without the constant presence of a caregiver. Such babbling seems to be motivated by the actual infant-produced sounds, which is a basic learning mechanism in *Listen and Babble*.

<sup>21</sup>Caregiver imitation: The caregiver responding to infant speech sounds with imitated, phonetically more correct, adult speech sounds.

Might there be any way of combining the strengths of both types of model (those based on how caregivers repond to infant actions and those based on the infant directly imitating perceptual targets)? The next and final section will attempt to answer that question.

## 5 Listen and Babble with Caregiver Imitation

In this last section, I would like to propose an extension of *Listen and Babble*, which:

- no longer separates perception and imitation into two different stages and
- attempts to solve the perceptual problem of the infant having to associate his own speech with that of his ambient speech (e.g. caregiver).

I suggest that the following model would:

- keep some of the advantages of *Listen and Babble* over other models of speech acquisition and
- explain the role of a caregiver *and* account for infants' babbling in the absence of the caregiver.

### A model including caregiver imitation

As in *Listen and Babble*, the infant's auditory system is trained (perceptual stage). However, this is done without including speech sounds produced by the infant vocal tract. This would be realistic: babbling infants simply don't hear articulate speech from (other) infants as young as themselves.

In *Listen and Babble*, only the infant himself listens to his speech sounds. I suggest that we include a caregiver who also listens to the infant's speech sounds. Caregiver imitation pairs every adequate infant sound with an adult speech sound of that very vowel. The infant would thus associate his own speech sounds, whenever good enough to be imitated, with adult perceptual categories.

The auditory system is then retrained (after each generation of speech samples), including those RL-agent's own sounds which were imitated by the caregiver, as well as the adult imitation itself. This auditory system returns confidences to the reinforcement learner. The infant auditory system is basically retrained after each generation of samples.

A question that remains: *how does the infant's auditory system retrain?*

As in the sensory training in *Listen and Babble*, a training set is produced (with ambient speech samples). This training set first consists of only non-infant speech samples (or e.g. samples from age 4 upward). Then, with every generation of self-produced speech, some new speech sounds (infant sounds, each with a corresponding imitated adult sound, and thus correctly labeled through association) are included in the training set. In each generation, new sounds replace 'older' sounds in the training set and in each generation, the auditory system is trained anew with the training set.

Over time, the quality of the infant's auditory system will increase for categorising its own sounds. I expect that the reinforcement learner will also yield increasingly good results as a consequence.

We could easily disable the caretaker at any time (or at certain intervals), thus returning to the original *Listen and Babble* model. How is learning influenced, when a caretaker is always present, constantly prompting the auditory system to be retrained? Or what happens if, after a while, the infant babbles along on his own for some time...?

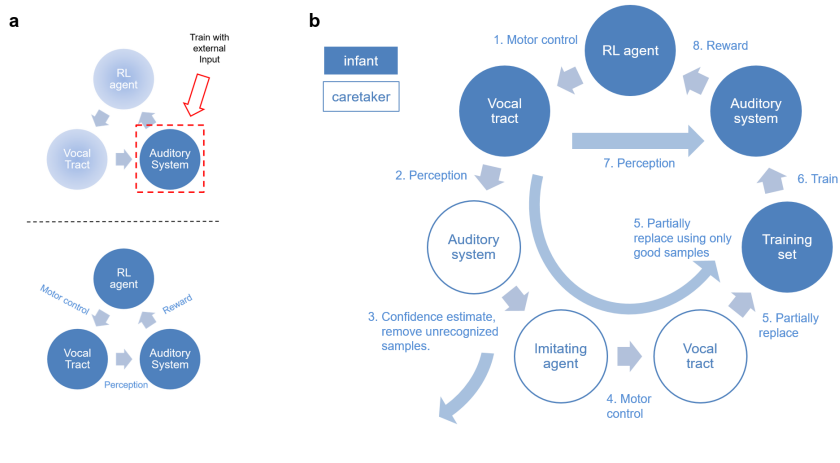


Figure 22: **a:** The *Listen and Babble* model. Two separate learning stages (target acquisition *top* and imitation *bottom*) **b:** A model including caregiver imitation.

In Figure 22, I schematically compare the original *Listen and Babble* model with the proposed extension. In step five, infant speech samples will be correctly labeled due to the fact that the caregiver has categorised them correctly. Providing an imitating sound for each sample, these pairs are simply given the same label before including them in the training set.

In Table 1, I sketched a possible algorithm for realising such a model. Some explanatory points ahead:

- The algorithm starts off with a fully trained caregiver auditory system (trained on infant + adult speech of all ages), an infant auditory system only trained on speech over a certain age. In a real sense the infant learner (realistically) starts off without any prior knowledge about:
  - how his own speech should sound (no perceptual target),
  - how to produce speech (the same as in *Listen and Babble*).
- The algorithm needs to work in parallel with:
  - $N_G$ - Generation size (could be 1000, with 100 CPUs!),
  - $N_I$  – A hand-set size of the 'best samples' in the current generation. (For example: if  $N_G = 1000$  and  $N_I = 10$ , the best 1% of the generation will be imitated by the caregiver).
- The caregiver agent could even get confidences in respect to all the classes, then label each sample with the class that yields maximum confidence. The infant will then receive imitated speech sounds from more then one class. This means, we could let the RL-Agent be intrinsically motivated and swap the current target on the go.
- Whenever the training set is “updated with the current generation’s samples”, only samples produced in this imitation are replaced. The training

set has a fixed set of ambient speech (from, say, age 4 upwards). After all, the infant would continue to perceive ambient speech. Only his own perceptual target (produced by the samples of age 1 in the training set) shifts.

An objection my readers might have:

*An auditory system that is (even at first) only trained on bad samples (those of maximal confidence will still not be good) cannot return reward signals accurate enough for the reinforcement learner to make any progress!*

That, I believe, is true if the auditory system is *only* trained on those 'bad samples'. But in fact, the (large portion) of training data (that of all older speakers) will be accurate. The additional infant's samples are only there to 'help bridge the gap' between its own sounds and those of, say, 4 year olds and upward. Having some (non-perfect) speech sounds of the infant himself included in the perceptual retraining might help the infant to generalise across speakers.

Who?	Com- ponent	Tasks	Resulting data (Size)
Infant	RL- Agent	<ul style="list-style-type: none"> <li>- choses <math>x\_mean</math> (from cma-es)</li> <li>- generates <math>N_G</math> offspring</li> <li>- return <code>infant_parameters</code></li> </ul>	<i>infant_parameters</i> ( $N_G$ )
	Vocal tract	<ul style="list-style-type: none"> <li>- synthesise infant speech using <code>infant_parameters</code></li> <li>- return <code>infant_speech_sounds</code> (*)</li> </ul>	<i>infant_parameters</i> ( $N_G$ ), <i>infant_speech_sounds</i> ( $N_G$ )
Care- giver	Auditory System	<ul style="list-style-type: none"> <li>- perception of <code>infant_speech_sounds</code></li> <li>- return caregiver confidences (**)</li> </ul>	<i>infant_parameters</i> ( $N_G$ ), <i>infant_speech_sounds</i> ( $N_G$ ), <i>cg_confidences</i> ( $N_G$ )
	Cgv.- Agent	<ul style="list-style-type: none"> <li>- sort data according to each sample's maximal confidence value (***)</li> <li>- keep first <math>N_I</math> of samples only, reject all others</li> <li>- imitate all <math>N_I</math> samples with their nearest class teacher parameters. These will be adult parameters.</li> <li>- return <code>adult_parameters</code> (<math>N_I</math>) for each of those <math>N_I</math> infant samples.</li> </ul>	<i>infant_parameters</i> ( $N_I$ ), <i>infant_speech_sounds</i> ( $N_I$ ), <i>cg_confidences</i> ( $N_I$ ), <i>adult_parameters</i> ( $N_I$ ), <i>adult_parameters</i> ( $N_I$ )
	Vocal tract	<ul style="list-style-type: none"> <li>- synthesise adult speech using <code>adult_parameters</code></li> <li>- return <code>adult_speech_sounds</code> (<math>N_I</math>)</li> </ul>	<i>infant_parameters</i> ( $N_I$ ), <i>infant_speech_sounds</i> ( $N_I$ ), <i>cg_confidences</i> ( $N_I$ ), <i>adult_parameters</i> ( $N_I$ ), <i>adult_parameters</i> ( $N_I$ ), <i>adult_speech_sounds</i> ( $N_I$ )
Infant	Auditory System	<ul style="list-style-type: none"> <li>- perceptual labeling: Each adult speech sound comes with the corresp. infant speech sound (pairs). Perception of each adult sound yields a label for the corresp. infant sound.</li> <li>- replace 'oldest' (****) samples in the TRAINING_SET (****) with the new labeled infant samples..</li> <li>- retrain reservoir with updated TRAINING_SET (which has now been updated with the 'good' samples)</li> <li>- (improved?) perception of <code>infant_speech_sounds</code></li> <li>- return <code>inf_confidences</code></li> </ul>	<i>inf_confidences</i> ( $N_I$ ) (Reset others)
	RL- Agent	<ul style="list-style-type: none"> <li>- compute reward of best sample</li> <li>- return to beginning (new <math>x\_mean</math>)</li> </ul>	

Table 1: A sketch of a possible algorithm to implement caregiver imitation.

(\*) *infant\_speech\_sounds* could be a list of paths to the wav files.

(\*\*) each sample will have a list of confidences for each class of the Auditory System

(\*\*\*) the maximum is taken over the classes.

(\*\*\*\*) TRAINING\_SET: initially consists of labeled samples from speakers over a certain age (e.g. 4 yrs). This global set of labeled speech sounds is regularly updated in each iteration and repeatedly used to train another reservoir for the infant's auditory system.

(\*\*\*\*\*) Replace only those samples produced by the infant learner (not ambient speech). Initially, no samples are replaced, however after mult. generations, older infant speech in the training set must be replaced in order to make progress.

## 6 For developers

### 6.1 Getting started

Each step in 'Results' is also documented in the project source code and can be reproduced by:

1. downloading *Listen and Babble* from [24],
2. installing needed dependencies,
3. executing the shell while having the relevant lines<sup>22</sup> in *control/get\_params.py* set as *True*,
4. checking result plots in the result directory or 'pickle-loading' the saved class in *data/classes/..pickle*.

### 6.2 Project structure

*Listen and Babble* is implemented using Python. I structured the code using *classes*, in which I grouped all functions belonging to one main stage of the project. Each stage of the project is executed from a project shell (in the main directory). The shell instantiates main classes, which in turn call functions from the functions class.

For an overview of the project directory, see Figure 23.

I advise first reading the documentation in the main directory [24], then shell and *get\_params* scripts, and after that to venture on to understand specific functions.

---

<sup>22</sup>Example:

In *control/get\_params.py*:

```
self.execute_main_script = {'ambient_speech':True,'hear':False,'learn':False}
self.do_setup = True
self.do_make_proto = True
self.do_setup_analysis = True
```

Then check results in *results/ambient\_speech/srange\_fm(2)/..*



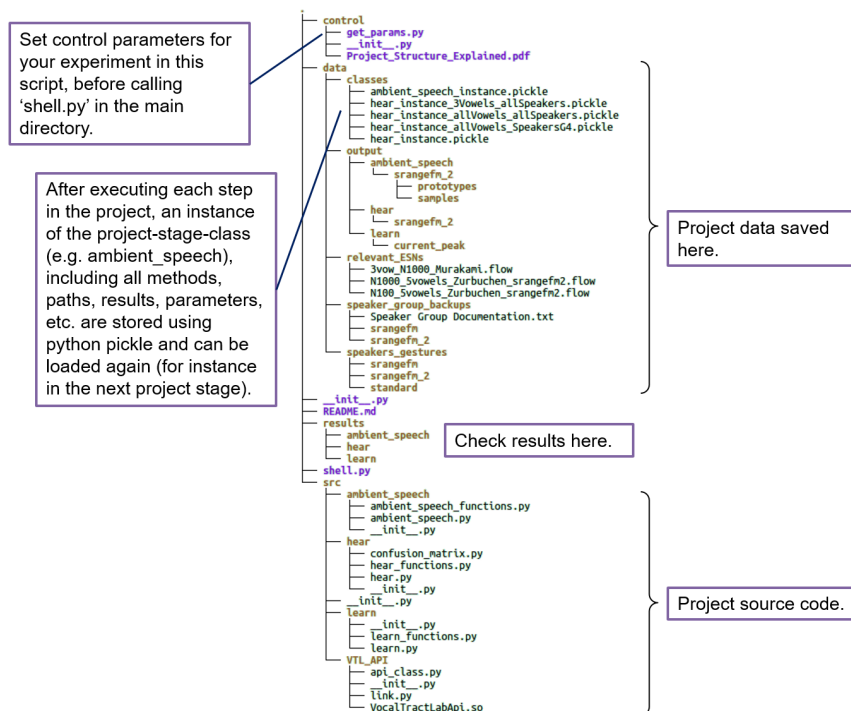


Figure 23: Project structure of my implementation of *Listen and Babble* [24].

# Appendix

# THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

© 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ		ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	ɓʼ Bilabial
ǃ (Post)alveolar	ɗ̥ Palatal	ɗʼ Dental/alveolar
ǂ Palatoalveolar	ɡ̊ Velar	ɡʼ Velar
ǁ Alveolar lateral	ɡ̊̚ Uvular	sʼ Alveolar fricative

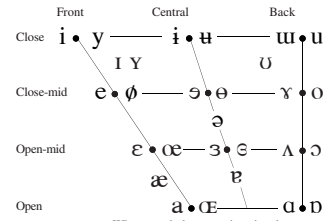
OTHER SYMBOLS

ɱ Voiceless labial-velar fricative	ɕ ʑ Alveolo-palatal fricatives
ɰ Voiced labial-velar approximant	ɺ Voiced alveolar lateral flap
ɥ Voiced labial-palatal approximant	ɮ Simultaneous ʃ and x
ʜ Voiceless epiglottal fricative	
ʕ Voiced epiglottal fricative	Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.
ʡ Epiglottal plosive	kp̚ ts̚

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ɲ̥̊

◌̥ Voiceless	◌̤ Breathy voiced	◌̦ Dental
◌̦ Voiced	◌̧ Creaky voiced	◌̨ Apical
◌̧ Aspirated	◌̨̥ Linguolabial	◌̩ Laminar
◌̨ More rounded	◌̩̥ Labialized	◌̪ Nasalized
◌̩ Less rounded	◌̪̥ Palatalized	◌̫ Nasal release
◌̪ Advanced	◌̫̥ Velarized	◌̬ Lateral release
◌̫ Retracted	◌̬̥ Pharyngealized	◌̭ No audible release
◌̬ Centralized	◌̭̥ Velarized or pharyngealized	
◌̭ Mid-centralized	◌̮ Raised	(◌̮ = voiced alveolar fricative)
◌̮ Syllabic	◌̯ Lowered	(◌̯ = voiced bilabial approximant)
◌̯ Non-syllabic	◌̰ Advanced Tongue Root	
◌̰ Rhoticity	◌̱ Retracted Tongue Root	

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

ˈ Primary stress
ˌ Secondary stress
ː Long
ˑ Half-long
˚ Extra-short
˘ Minor (foot) group
˙ Major (intonation) group
˙ Syllable break .i.ækt
˘ Linking (absence of a break)

TONES AND WORD ACCENTS	
LEVEL	CONTOUR
˥ Extra high	˥ or ˩ Rising
˨ Extra low	˥ or ˩ Falling
˨̥ High rising	˥ or ˩ High rising
˨̥̥ Low rising	˥ or ˩ Low rising
˨̥̥̥ Extra low rising	˥ or ˩ Extra low rising
˩ Downstep	˥ or ˩ Global rise
˩ Upstep	˥ or ˩ Global fall

## References

- [1] Murakami, B. Kröger, P. Birkholz, J. Triesch. Seeing [u] aids vocal learning: Babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing. *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. 2015.
- [2] Murakami, Listen and Babble – A model of Vowel Acquisition Based on Imitation Learning. *Thesis*. 2014.
- [3] <http://ieeexplore.ieee.org/document/7346142/>
- [4] P. Kuhl, F. Tsao, H. Liu, Y. Zhang, and B. De Boer. Language/culture/mind/brain. progress at the margins between disciplines. *Annals of the New York Academy of Sciences*, 935, 2001.
- [5] P. Kuhl. Learning and representation in speech and language. *Curr. Opin. Neurobiol.* 4:812-822, 1994.
- [6] P. Kuhl. A new view of language acquisition. *Proc. Natl. Acad. Sci. USA* 97(22): 11850-11857.
- [7] P. Kuhl. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5, 2004.
- [8] Boersma, Paul & Weenink, David (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.19, retrieved 01.03.2016 from <http://www.praat.org/>
- [9] 'Phoneme.' Merriam-Webster.com. <http://www.merriam-webster.com> (18.07.2016).
- [10] 'Pronunciation of English ⟨th⟩.' Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 07.06.20016. Web. 18.07.2016.
- [11] P. Birkholz. Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis. *PLoS ONE*, 8, 2013.
- [12] K. Johnson. Speaker normalization in speech perception. In D. Pisoni and R. Remez, editors, *The Handbook of Speech Perception*, section 15. 2008.
- [13] P.D. Eimas, J.L. Miller. Contextual effects in infant speech perception. *Science* 209. 1140-1141. 1980.
- [14] Miller, J.L. & Liberman, A.M. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Percept. Psychophys.* 25, 457-465. 1979.
- [15] Conversational interfaces: advances and challenges. *Proc. IEEE* 88 1166-1180. 2000.
- [16] Kuhl, P.K. Perception of auditory equivalence classes for speech in early infancy. *Infant Behav. Dev.* 7, 263-285. 1983.

- [17] N. Mesgarani, C. Cheung, K. Johnson, E. F. Chang. Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science* 343. 2014.
- [18] Jones SS. The development of imitation in infancy. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 364(1528):2325-2335. 2009.
- [19] A.N. Meltzoff & M.K. Moore. Imitation of facial and manual expressions by human neonates. *Science*. 198, 75-78. 1977.
- [20] E.A. Simpson, L. Murray, A. Paukner, P.F. Ferrari. The mirror neuron system as revealed through neonatal imitation: presence from birth, predictive power and evidence of plasticity. *Phil. Trans. R. Soc. B* 369. 2014.
- [21] B. Mampe, A.D. Friederici, A. Christophe, K. Wermke. Newborns' Cry Melody Is Shaped by Their Native Language. *Current Biology* Volume 19, Issue 23, p1994-1997. 15.12.2009
- [22] P. Roach. Phon2. <http://www.personal.rdg.ac.uk/~llsroach/phon2/artic-basics.htm>. *University of Reading*. Web. 20.07.2016
- [23] S. Lee, A. Potamianos, Sh. Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Am.* 105 (3), 1999
- [24] P. Zurbuchen. ListenAndBabble repository. <https://github.com/PhilJZ/ListenAndBabble>. 10. September, 2016.
- [25] J. Hillenbrand, L. Getty, M. Clark, K. Wheeler. Acoustic Characteristics of American English Vowels. *JASA* 97(5) 1995.
- [26] J. P. Burg, 'Maximum Entropy Spectral Analysis,' PhD thesis, Stanford University, 1975.
- [27] A. Kabir, J. Barker, M. Giurgiu. Robust Formant Estimation: Increasing the Reliability by Comparison among Three Methods. *Proc. of the Intern. Conf. on Circ., Syst., Sig.* Received: 12.08.2010
- [28] R.D. Kent. The uniqueness of speech among motor systems. *Clinical Linguistics & Phonetics* Volume 18, Pages 495-505. September 2004.
- [29] Owens, R.E. (2005). *Language Development: An Introduction*. Boston: Pearson. pp. 125-136.
- [30] Harley, Trevor (2001). *The Psychology of Language* 2Ed. New York: Psychology Press. ISBN 0-86377-867-4.
- [31] Werker, Janet F; Tees, Richard C. (1999). 'Influences on infant speech processing: Toward a new synthesis.'. *Annual Review of Psychology* 50: 509-535. doi:10.1146/annurev.psych.50.1.509.
- [32] B. Tucker. Labeled vocal tract. <http://aphl.artsrn.ualberta.ca/?p=247>, 2013. [Online; accessed 30-July-2016].
- [33] I.R. Titze. Physiologic and acoustic differences between male and female voices. *J. Acoust. Soc. Am.* 85 (4). April 1989.

- [34] H.Hanson, E.Chuang. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *J. Acoust. Soc. Am.* 106 (2). August 1999.
- [35] S.J. Pawlby. A study of imitative interaction between mothers and their infants (Ph.D.). University of Nottingham. 1977a.
- [36] Reward signals. [http://www.scholarpedia.org/article/Reward\\_signals](http://www.scholarpedia.org/article/Reward_signals). [Online; accessed 2.10.2016].
- [37] Miller, R.L. Auditory tests with synthetic vowels. *J. Acoust. Soc. Am.* 25 , 114-121. 1953
- [38] Slawson, A.W. Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *J. Acoust. Soc. Am.* 43 , 87-101. 1968.
- [39] Fujisaki, H. and Kawashima, T. (1968) The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics* AU-16 , 73-77.
- [40] E. Lopez-Poveda and R. Meddis. A human nonlinear cochlear filterbank. *J. Acoust. Soc. Am.* 110, 2001.
- [41] B. Fontaine, D. Goodman, V. Benichoux and R. Brette. Brian hears: online auditory processing using vectorization over channels. *frontiers in Neuroinformatics*, 5, 2011.
- [42] D. Goodman and R. Brette. Brian: a simulator for spiking neural networks in Python. *frontiers in Neuroinformatics*, 2, 2008.
- [43] H. Jaeger. Echo state network. *Scholarpedia*, 2007.
- [44] 'CMA-ES' Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 07.06.20016. Web. 27.08.2016.
- [45] Lukoševičius M. A Practical Guide to Applying Echo State Networks. In: G. Montavon, G. B. Orr, and K.-R. Müller (eds.) *Neural Networks: Tricks of the Trade*, 2nd ed. Springer LNCS 7700, 659-686. 2012
- [46] Voice changes throughout life. *National Center for Voice and Speech*. <http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/changes.html> [Online; accessed 24.10.2016]
- [47] Zahner, K., Schönhuber, M., Grijzenhout, J. & Braun, B. (2016). Konstanz prosodically annotated infant-directed speech corpus (KIDS Corpus). Proceedings of the 8th International Conference on Speech Prosody. Boston, USA.
- [48] Miura K, Yoshikawa Y, Asada M (2007) Unconscious anchoring in maternal imitation that helps finding the correspondence of caregiver's vowel categories. *Advanced Robotics* 21 (13): 1583–1600. doi: 10.1109/roman.2007.4415121
- [49] Y. Yoshikawa, M. Asada, K. Hosoda, and J. Koga. A constructivist approach to infants' vowel acquisition through mother–infant interaction. *Connection Science* Vol. 15 , Iss. 4. 2003.

- [50] Ian S. Howard, Piers Messum. Modeling the Development of Pronunciation in Infant Speech. *Acquisition85 Motor Control*, 15, 85-117. 2011.
- [51] Ian S. Howard, Piers Messum. Learning to Pronounce First Words in Three Languages: An Investigation of Caregiver and Infant Behavior Using a Computational Model of an Infant. PLoS ONE 9(10): e110334. doi:10.1371/journal.pone.0110334. 2014.
- [52] Messum, P., & Howard, I. S. Creating the cognitive form of phonological units: The.... *Journal of Phonetics*. (2015)

Erklärung nach § 30 (12) Ordnung für den Bachelor- und den Masterstudien-  
gang:

Hiermit erkläre ich, dass ich die Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst habe. Alle Stellen der Arbeit, die wörtlich oder sinngemäß aus Veröffentlichungen oder aus anderen fremden Texten entnommen wurden, sind von mir als solche kenntlich gemacht worden. Ferner erkläre ich, dass die Arbeit nicht – auch nicht auszugsweise – für eine andere Prüfung verwendet wurde.

Frankfurt, den